Recommender Systems Leveraging Multimedia Content

YASHAR DELDJOO, Polytechnic University of Bari, Italy
MARKUS SCHEDL, Johannes Kepler University Linz and Linz Institute of Technology, Austria
PAOLO CREMONESI, Politecnico di Milano, Italy
GABRIELLA PASI, University of Milano-Bicocca, Italy

Recommender systems have become a popular and effective means to manage the ever-increasing amount of multimedia content available today and to help users discover interesting new items. Today's recommender systems suggest items of various media types, including audio, text, visual (images), and videos. In fact, scientific research related to the analysis of multimedia content has made possible effective content-based recommender systems capable of suggesting items based on an analysis of the features extracted from the item itself. The aim of this survey is to present a thorough review of the state-of-the-art of recommender systems that leverage multimedia content, by classifying the reviewed papers with respect to their media type, the techniques employed to extract and represent their content features, and the recommendation algorithm. Moreover, for each media type, we discuss various domains in which multimedia content plays a key role in human decision-making and is therefore considered in the recommendation process. Examples of the identified domains include fashion, tourism, food, media streaming, and e-commerce.

CCS Concepts: • **Information systems** → **Recommender systems**; **Multimedia and multimodal retrieval**; • **Human-centered computing** → **User models**;

Additional Key Words and Phrases: Content-based recommender systems, multimedia, machine learning, deep learning, signal processing, audio, music, image, video, fashion, food, e-commerce, tourism, social media

ACM Reference format:
Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2020. Recommender Systems Leveraging Multimedia Content. *ACM Comput. Surv.* 53, 5, Article 106 (September 2020), 38 pages.
https://doi.org/10.1145/3407190

1 INTRODUCTION

Information overload has become a serious issue of the modern society. As a remedy, recommender systems (RS)[1] have emerged as a paradigm of information push. They aim to mitigate the negative impacts of the over-choice burdened on users, which can result in poor decision-making,

[1]Please note that we provide a list of abbreviations used in the survey as an appendix.

Authors' addresses: Y. Deldjoo (corresponding author), Polytechnic University of Bari, Via Orabona 4, Bari, 70125, Italy; emails: yashar.deldjoo@poliba.it, deldjooy@acm.org; M. Schedl, Johannes Kepler University Linz and Linz Institute of Technology, Altenberger Straße 69, Linz, 4040, Austria; email: markus.schedl@jku.at; P. Cremonesi, Politecnico di Milano, p.zza Leonardo da Vinci, 32, Milan, 20133, Italy; email: paolo.cremonesi@polimi.it; G. Pasi, University of Milano-Bicocca, Viale Sarca 336, Milan, 20126, Italy; email: gabriella.pasi@unimib.it.

0360-0300/2020/09-ART106 $15.00
https://doi.org/10.1145/3407190

(emotional) exhaustion, and loss of time. RS are data-driven algorithms whose goal is to present to users a selected and personalized subset of items from a huge set of distinct candidate items. As such, RS reduce the effects of information overload and help users to more easily identify what they could be interested in, therefore increasing the quality of their decision-making process. A massive amount of research has been devoted to improve RS, not least because of initiatives such as the Netflix Prize,[2] driven by commercial interests [Koren et al. 2009].

Algorithms for RS can be broadly classified into collaborative filtering (CF), context-aware (CA), content-based filtering (CBF), and hybrid.

- Collaborative Filtering (CF) models rely on historical behavioral data of a community of users, either past interactions with items (e.g., previous clicks, check-ins, purchases) or explicit preferences (e.g., rating scores), to identify preference patterns and predict the utility of unknown items to an active user [Koren and Bell 2015]. CF models can be classified into *memory-based* and *model-based* approaches. Memory-based approaches directly work with the dataset of interactions, while model-based approaches assume an underlying model that explains the user-item interactions and try to discover it to make recommendations [McFee et al. 2012; Yuan et al. 2016]. Model-based approaches can be further categorized into *linear methods*, such as matrix factorization (MF) and graph-based methods, and *nonlinear methods*, such as the ones based on deep neural networks (DNN).
- Context-aware (CA) models enhance CF by including contextual information into the model (e.g., time or location) to capture the current need of the user [Aggarwal 2016b].
- Content-based filtering (CBF) models require as input the target user's behavioral data (but not those of non-target users') together with item content information, typically represented as real-values vectors [Hu et al. 2008]. To give some examples in the multimedia domain, item properties can be words or concepts in a text, colors in an image, amount of motion in a movie, or rhythm in a music piece.
- Hybrid recommender systems combine two or more of the previous approaches, using a fusion approach [Burke 2002; Çano and Morisio 2019].

Most CBF methods (or hybrids thereof) receive text as the only input media type (e.g., the description of a product, the plot of a movie, the synopsis of a book). Fewer CBF approaches also process aural or visual content, together with text, to generate better quality recommendations. The purpose of this survey is to provide a comprehensive review of RS that incorporate multimedia (MM) content in the design space of the recommendation model. As such, the recommendation methods that are of interest in this survey are either pure CBF models or hybrid systems that combine CBF with CF or CA approaches. In this context, multimedia content (i.e., combinations of audio, visual, text) can be used in two recommendation scenarios to recommend:

(1) a particular media item to a user—for example, a music track or a movie; or
(2) a non-media item by leveraging multimedia content related to that item—for example, to recommend clothes based on the visual appearance of respective photos.

In this survey, both scenarios are addressed. We analyze RS that exploit at least one of aural or visual modalities to recommend either media items (e.g., , music, pictures, movies, user-generated videos) or non-media items that have media-related attributes (e.g., a fashion item described with pictures and text). We highlight that pure text-based CBF models have been studied for long in the RS community. We therefore do not review pure textual content-based recommenders [Aggarwal 2016a; de Gemmis et al. 2015], but we include approaches where text is not the unique

[2] http://www.netflixprize.com.

kind of item description. The motivation is to analyze the impact of audio and visual stimuli on user preferences.

Terminology: Since the term "multimedia." is not unambiguously defined in literature, it is vital to clarify the following notions, which we use throughout the survey:

- We use the term "modality" to refer to the information extracted from different signal channels. We consider three basic modalities: aural, visual, and textual (A, V, T).
- We use the term "media item" to refer to an item composed of one or several (A, V, T) modalities.

A, V, and T serve as basic atoms describing various media items. To give a few examples of media items (and constituting modalities): picture (V), pop song (A + T), piano sonata (A), silent movie (V), regular movie (A + V + T), news article with surrounding images (A + T). Based on this definition, we can consider all examples in which a single modality is used as *unimodal* or *atomic* media items; when more than one modality is included as *multimodal* or *composite*.[3]

The practical outcome of the present survey is:

- It highlights more than 10 domains in which multimedia content (processing) is used to solve a real-world recommendation problem. Examples of these domains include: music, video, fashion, food, e-commerce, social media, cultural heritage, and tourism.
- It underscores the importance of visual and aural properties of items interacted with in modeling a user's profile.
- It unifies the advances made in the communities of MM and RS by presenting a framework of RS leveraging MM content, which we use to categorize the reviewed research works (cf. Section 1.1).
- It highlights that multimedia content can be useful to recommend items that are not necessarily media types but may also be generic items (e.g., fashion items or food).
- Additionally, we created an open-source repository that includes all reviewed articles categorized with respect to target domain.[4] We hope this repository will facilitate benchmarking multimedia-related projects in the RS field by providing links to respective code and datasets.

1.1 The Pipeline of Recommender Systems Leveraging Multimedia Content

A recommender system leveraging multimedia content generally follows the steps sketched in Figure 1 and described in the following. Please note that only important steps in the demonstrated pipeline are used for categorization in Tables 1, 2, 3, and 4.

1. Segmentation: The goal of segmentation is to partition the media signal into segments that are *homogeneous* in some feature space. The motivation is efficiency (e.g., a video can contain thousands of similar frames) and improving informativeness of the final descriptor. In the case of audio, segmentation is achieved via *framing*, e.g., splitting the raw audio signal into frames covering a number of samples that equal tens or hundreds of milliseconds. In case of images, segmentation is performed via *clustering image regions*, e.g., with respect to color, a process known as spatial segmentation. For videos, the segmentation can be done both in space and time. In case of text, segmentation is commonly realized via *tokenization*, i.e., splitting a text into words or n-grams (sequences of n words). Please note that segmentation is an optional step and not always performed for all media types, e.g., also the raw pixel values of an image can be fed into a DNN.

[3]Our definition is consistent with the literature of multimedia information retrieval [Mayer 2005; Stanciulescu 2008]. Sometimes, the terms "sensory modality level." and "physical level." are used to refer to the same concepts.
[4]https://github.com/yasdel/mmrecsys_survey20.

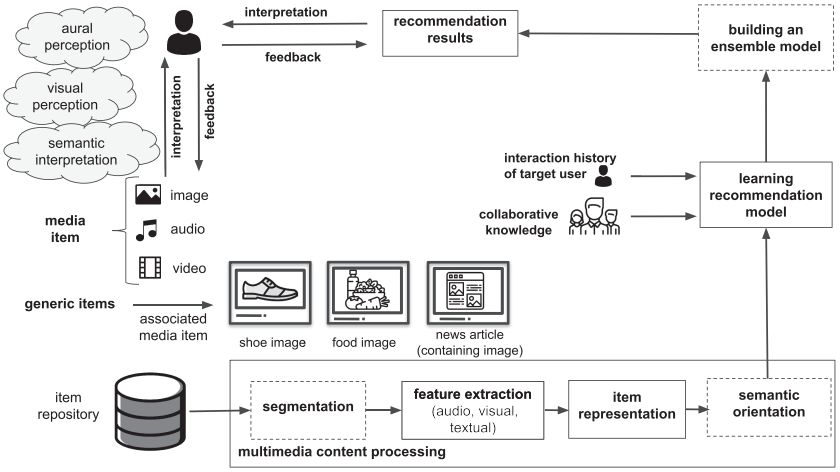

Fig. 1. Generic framework for recommender systems leveraging multimedia content.

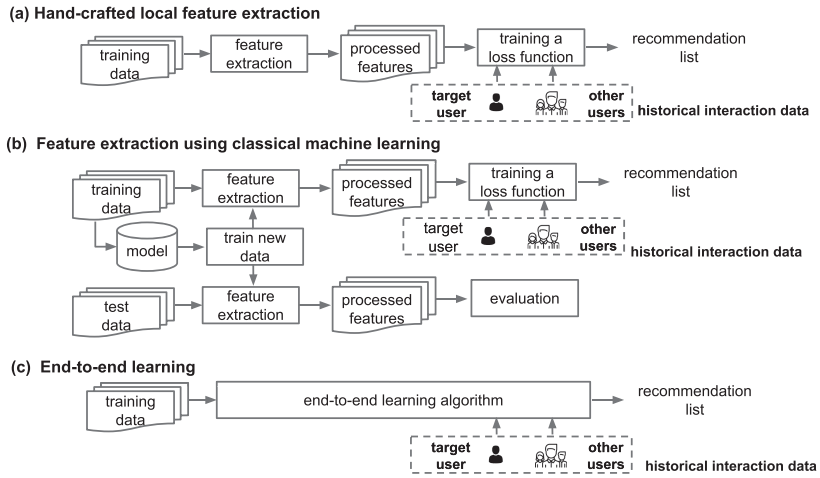

Fig. 2. Different paradigms of feature extraction for recommendation task leveraging multimedia content: (a) handcrafted feature extraction, (b) machine-learned feature extraction, (c) end-to-end learning.

2. Feature Extraction: The goal of feature extraction is to describe the content of a media item in a low-dimensional and descriptive way, so it can be exploited in subsequent processing steps. From an algorithm point-of-view, we can identify three paradigms of feature extraction as illustrated in Figure 2:

- *Handcrafted features:* Before the advent of representation learning, handcrafted features, designed by domain experts, were predominantly used in all tasks related to multimedia processing.
  While, on one hand, the handcrafted features can introduce the risk that some interesting but non-intuitive phenomena may be neglected in the definition of features; on the other hand, their advantage is that they are often *semantically interpretable*, making them suitable for explanation purposes in RS.

- *Feature extraction using classical machine learning:* Such methods involve training a machine learning model on a specific dataset and represent new items (from a comparable distribution) by applying the learned model. Examples of this type of feature include i-vectors [Dehak et al. 2011] in the audio and Fisher vectors in the image domain [Perronnin and Dance 2007].
- *End-to-end (E2E) learning:* As a specific class of machine learning, the motivation behind E2E learning is to eliminate the need for handcrafted heuristics. E2E unifies feature extraction and model learning into a single step and has been highly popularized in the context of DNNs. In a deep E2E system, the system is only provided with the raw data (e.g., images) and the target user's feedback (e.g., user's item ratings), and the network learns parameters across multiple layers (> 3) of a DNN that can map the raw data to output directly, without the need to extract intermediate features.

With the exception of E2E learning, which does not produce feature vectors, the results of feature extraction are descriptors representing item content. We classify them into:

(1) *Basic signal-level features*: For audio (Table 1), these include *energy*, *spectral*, *timbre*, and *beat*-related features. For images (Tables 2 and 3), they include *color*, *texture*, *edge*, and *shape*.
(2) *Specific feature extraction method*: Beyond basic signal-level features, we distinguish a number of popular audio processing and computer vision algorithms used for feature extraction.
    - Aural features: The common methods to extract features from raw audio signals, as shown in Table 1, include: short-time Fourier transform (*STFT*) [Allen 1977], Mel frequency cepstral coefficients (*MFCC*) [Logan 2000], and automatic feature learning based on a convolutional neural network (*CNN*) fed with the spectrogram [Oramas et al. 2017; Schlüter 2016].
    - Visual features: The most common methods to extract features from image content (Tables 2 and 3) include: speeded up robust features (*SURF*) [Bay et al. 2006], scale invariant feature transform (*SIFT*) [Lowe 2004], local binary patterns (*LBP*) [Ojala et al. 2002], Gabor filters [Manjunath and Ma 1996], discrete Fourier transform (*DFT*), histogram of oriented gradients (*HOG*) [Dalal and Triggs 2005], and *CNN* [Krizhevsky et al. 2012].

3. Item Representation: To obtain a standardized representation of each item e.g., fixed-length feature vector, extracted features typically undergo an additional processing step, such as bag of (visual) words (BOWs) modeling, feature aggregation (e.g., temporal aggregation of audio features), or dimensionality reduction. In text, the most popular item representation technique is the *BOW* model, where the features employed to formally represent a text are terms (words) or n-grams. Similarly, such a BOW model can be adapted to other domains. For images, salient points (e.g., detected by SIFT) can be clustered using *k-means*, such that each cluster center represents a "visual word." The image is then represented by a weight vector over these visual words. In audio, the same approach can be applied using MFCC vectors as "audio words." As an alternative to k-means, features can be aggregated by Gaussian mixture models (*GMM*), which describe all feature vectors of a given item as a fixed number of Gaussian distributions (as means, covariances, and mixture weights). Vector quantization (*VQ*) is an efficient approach to represent feature vectors using some statistics over a finite set of prototype vectors that form a codebook.

In addition to simple BOW models, terms are often weighted using the *tf-idf* weighting function or the Okapi BM25 term weighting, e.g., Baeza-Yates and Ribeiro-Neto [2011]. More recently, text representations that consider the context of words (i.e., terms that occur in similar contexts are

semantically similar) have become popular. Word embeddings are then dense words representations in which similar words have a similar context. *Word2vec* [Mikolov et al. 2013a] and *Glove* [Pennington et al. 2014] are two popular algorithms aimed at producing such word embeddings. For what concerns data fusion, we consider three types of fusion: *early fusion*, *late fusion*, and *others*. In early fusion, features are combined before feeding them into the prediction system; for example, by concatenation of feature vectors or more advanced methods such as canonical correlation analysis (CCA) [Deldjoo et al. 2019]. In late fusion, the outputs of different prediction systems are combined; for instance, consider rank aggregation method based on the Borda count [Deldjoo et al. 2018a] and priority-aware fusion [Du et al. 2020] approach used for multimodal data fusion in MMRS. Other fusion techniques that cannot easily be categorized into early or late include: re-ranking methods [Andjelkovic et al. 2019], combining features in a latent space [Liu et al. 2014] or at intermediate similarity-level [Bartolini et al. 2013; Canini et al. 2013], neural approaches [Ma et al. 2018], kernel-based approaches [Deng et al. 2013], optimization-based approaches [Cui et al. 2014], and graph-based fusion [Cui et al. 2010; Farseev et al. 2017].

4. Semantic Orientation: Depending on the application in question, the features extracted from the media items may be *semantically oriented* into prediction of particular target classes in the expectation that this will eventually have a positive impact on the final quality of recommendation. For instance, features extracted may be oriented into prediction of some predefined emotion or genre classes, human-generated tags (auto-tagger), or some latent factor features of CF models. We also consider approaches relying on DNNs pre-trained for tasks other than RS. In Tables 1 to 3, we use the notation sig−>tag (i.e., MM signal is used to predict tags), sig−>CF-LF (i.e., MM signal is used to predict latent factors for CF), and sig−>NN (i.e., MM signals are projected into a latent space represented by a neural network) to show these kinds of features semantic orientation.

5. Learning Recommendation Model: Let $\mathcal{U}$ and $\mathcal{I}$ denote a set of real users and items, and $g : \mathcal{U} \times \mathcal{I} \rightarrow \mathbb{R}$ a utility function. A recommendation problem is defined as:

$$\forall u \in \mathcal{U}, i'_u = \underset{i \in \mathcal{I}}{\operatorname{argmax}}\ g(u, i), \tag{1}$$

where $g(u, i)$ is the *estimated utility* of item i for user u, based on which the items are ranked and i'_u is an item not consumed by u before.

Definition (Recommendation by exploiting MM content). Given $\mathcal{U}$ and $\mathcal{I}$, assuming that each item $i \in \mathcal{I}$ can be described in terms of one or a combination of aural (A), visual (V), and textual (T) modalities, the **recommendation problem by leveraging MM content** can be formulated as:

$$\forall u \in \mathcal{U}, i'_u = \underset{i \in \mathcal{I}}{\operatorname{argmax}}\ f(g_d(u, i)), \quad d \in \{V, A, T\} \tag{2}$$

in which $g_d(u, i)$ is the utility computed across several modalities $d \in \{V, A, T\}$, where f is an aggregation operator of the utility computed on each modality [Deldjoo et al. 2018d].

In overview tables 1, 2, 3, and 4, we classify recommendation models according to their class (CBF, CF, CA), specific type (memory-based, latent factor model (LFM), graph-based, and deep neural network models), and finally whether they represent an E2E system. When dealing with CA systems, we distinguish two types of contexts: multimedia and situational. The former is when the system requires a user to proactively provide a multimedia item to activate the recommendation process (e.g., recommending accessories that complement an item the user is looking at). In contrast, situational context refers to factors such as time or location. We use the mark $\surd^*$ throughout the tables to highlight *multimedia context* and *E2E systems*.

1.2 Strategy for Literature Search

To identify the publications relevant for this survey, we adopted a multi-level search strategy. We started by finding relevant research works from top conferences in the fields of RS and MM, i.e., the ACM Conference on Recommender Systems and ACM Conference on Multimedia, respectively, and collected a majority of the publications that address topics within the scope of the survey. Having in mind that many other venues on (multimedia) information retrieval and multimedia signal processing also publish relevant works, we also gathered a number of related publications by searching on Google Scholar and filtering for the name of the following journals and organization conferences: *IEEE Transactions on Multimedia; Journal of Multimedia Tools and Applications; ACM Transactions on Intelligent Systems and Technology; IEEE Transactions on Audio, Speech, and Language Processing;* ACM Special Interest Group on Information Retrieval conferences; and International Society for Music Information Retrieval conferences. Last, we performed some free search in Google Scholar and obtained a number of papers from other venues. We focused on conference proceedings and journals, to a much lesser extent on workshop publications. After having read the publications identified as explained above, we looked for the closest work to each publication, usually cited in the target publication and sometimes by using the "Related Work" option in Google Scholar. This way, we obtained a second list of publications. While we are quite sure that we did not find *all* publications that address a recommendation problem by leveraging multimedia content, we are confident that we could identify a vast majority of relevant publications.

1.3 Survey Context and Related Surveys

While there exist several survey articles on RS topics (e.g., Bobadilla et al. [2013], Ekstrand et al. [2011], Holey and Prabhune [2014], Park et al. [2012], Rafsanjani et al. [2013], Shi et al. [2014]), to the best of our knowledge, none of them focuses on recommendation in the multimedia domain based on content features. In contrast, the current survey at hand provides a comprehensive study of algorithms and systems that exploit multimedia content information and possibly uncover relationships between modalities. This can, in turn, provide RS with a rich and diverse source of information that can be relevant to the recommendation task.

The survey at hand is, nevertheless, related to other literature reviews on recommender systems, which take different perspectives. In terms of recommendation domain, there exist surveys that focus on e-commerce [Wei et al. 2007] or music [Bonnin and Jannach 2014; Knees and Schedl 2013], or only a specific use case, e.g., MMRS for mobile environments in smart communities [Xia et al. 2013]. Technology-wise, recent surveys on adversarial machine learning [Deldjoo et al. 2020c] and deep learning [Zhang et al. 2019] in RS present recent neural breakthroughs to solve hard tasks in MMRS. Fashion item generation or complementary recommendation are examples of such non-trivial tasks that have been addressed by using generative adversarial networks (GANs) or Siamese networks reviewed in these surveys.

Also, we would like to point to the recent line of research on interpretability and explainability of multimedia recommendations, among others in the fashion domain. While we acknowledge this important trend, surveying these topics is beyond the scope of the present article, but a respective survey can be found in Zhang and Chen [2020].

2 RECOMMENDER SYSTEMS LEVERAGING MULTIMEDIA CONTENT

In this main section of the survey, an overview of research on RS using multimedia content is provided. The goal of our work is to highlight the advances in both recommender systems and multimedia, bridging the two corresponding communities. Thereby, our objective is to show various domains and tasks in which audio, image, and video processing have been useful successfully

for recommendation tasks. We therefore organize the section according to the three most common media types used in addition to text: *audio*, *image*, and *video*. For each media type, we further classify RS based on the family of recommended items:

- Audio-based recommendation (Section 2.1, Table 1)
  - —Sound recommendation
  - —Music recommendation
- Image-based recommendation (Section 2.2)
  - —Image recommendation (Table 2)
  - —Product recommendation based on image-related attributes (Table 3)
- Video-based recommendation (Section 2.3, Table 4)
  - —Movie recommendation
  - —User-generated video recommendation

To further structure our review, we use `typewriter font` to highlight major `tasks` addressed in the discussed research works and *italics* to point to *important aspects* in the paper (e.g., methods, components thereof, and other noteworthy aspects).

2.1 Audio Recommendation

Recommender systems whose target items are given by pieces of audio content can be categorized according to the type of recommended item into **speech recommenders** (e.g., newscasts or podcasts), **sound recommenders** (e.g., sound effects for movies), and **music recommenders** (e.g., songs or artists). We survey recent work on speech recommendation in Section 2.1.1, on sound recommendation in Section 2.1.2, and on music recommendation in Section 2.1.3; the latter is, by far, the largest research topic in the area of audio recommendation.

2.1.1 Speech Recommendation. Unfortunately, existing approaches tailored to content-based speech recommendation (podcasts or newscasts) are not yet mature enough or available in a considerable number. For these reasons, we refrain from providing a detailed review of the few and specific approaches, such as those presented in Xing et al. [2016] for `Chinese podcast recommendation` based on textual information associated with the audio items, or Yang et al. [2018], which focuses on the evaluation of `conversational interfaces` for podcast recommendation.

2.1.2 Sound Recommendation. The topic of sound RS has not been addressed in the literature as prominently as, for instance, music RS. Nevertheless, systems that are capable of recommending sounds (e.g., for movies or producing electronic music) have their niche in RS research.

In Ostuni et al. [2015], the authors propose a `sound recommender system` that leverages user-generated tags and sound descriptions—which they enrich by *entity extraction*—that are later linked to external knowledge bases (KB) such as WordNet[5] and DBpedia.[6] The enriched data are then combined using a domain-specific tagging ontology represented by a *knowledge graph* (KG). Preliminary validation of the proposed system using Freesound[7] as a case study shows that such KBs can improve sound recommendation accuracy over several state-of-the-art CF baselines (such as BPR and SLIM).

In Oramas et al. [2016], the authors present a hybrid approach for `recommending sounds to music producers` by integrating *knowledge graphs* (KG) with information extracted from

[5] https://wordnet.princeton.edu.
[6] https://wiki.dbpedia.org.
[7] Freesound (https://www.freesound.org) is a collaborative sound repository where users can share recorded sounds.

documents describing audio items. While the authors do not mention the exact acoustic features used, they seem to cover the majority of descriptors available in Essentia.[8] The proposed approach consists of two steps: (1) item data enrichment and linkage to KG and (2) building of a graph-based recommender system on top of the KG. In the former step, the authors' approach relies on *entity linking*, which is performed by extracting semantic entities from item textual description and linking them to WordNet and DBpedia for gathering additional knowledge. Information extracted from these complementary resources is represented with a new KG. This graph is later merged with collaborative, implicit feedback to provide recommendations. The authors also propose several graph-embedding techniques to represent a KG as a (linear) feature combination. They validate their proposed solutions for two different use cases: (1) song and artist recommendation targeted to music consumers and (2) sound recommendation targeted to music producers. This research demonstrates that KBs such as DBpedia or WordNet help in building recommender systems for both music producers and consumers. Furthermore, entity linking on KG allows to boost recommendation performance in terms of novelty and diversity.

To save time and effort in `accessing sound repositories`, in Smith et al. [2019] the authors present a hybrid recommender for music producers, which combines CF and acoustic CBF to suggest sounds for users of the EarSketch sound browser.[9] EarSketch introduces a novel strategy for teaching computer science ideas through algorithmic composition of music. Learners can write up Python or JavaScript codes to control sound examples while understanding fundamental programming ideas, such as loops, lists, conditions, and functions [Mahadevan et al. 2015; Smith et al. 2019]. A shortcoming of EarSketch is that the majority of scripts written by users exploit only a small portion of the sound library. By integrating the proposed sound recommender in EarSketch, the system's basic search and filtering functionality is extended to recommend from the sound library's relevant sounds, thereby fostering novel, diverse, and serendipitous encounters. The system uses as input one or more sounds of a work-in-progress script. A recommendation score is computed for each candidate sound. An *item-based CF* component is used to add relevance to the generated recommendations, while the *audio-based CBF* components are utilized to increase the novelty of recommendations. The authors indicate the proposed sound recommender can inspire users' musical creativity and stimulate personal experience by exploring new sound libraries.

2.1.3 Music Recommendation. The majority of research on this topic assumes a passive user, i.e., the user is not required to activate the RS, via a query, rather recommendations are generated automatically. In contrast, some systems aim at recommending music for specific scenarios; for example, to serve as background music for a user-generated video or slideshow. Others aim at recommending coherent sequences of songs, i.e., playlists. Hence, we structure the following review into three categories: `query-free music recommendation` (the standard recommendation scenario), `music recommendation to enrich other media types` (taking an additional query as input), and `automatic playlist generation` (sequence-aware recommendation).

Query-free music recommendation: Content-based music recommendation in its classical flavor, i.e., recommending items/songs to a user based on user's interest on item on descriptions, has been researched since the mid 2000s. In the following, we focus on most recent state-of-the-art research, but we start with a discussion of noteworthy early works.

In Yoshii et al. [2007, 2008], the authors propose a recommendation strategy to solve the `accuracy-diversity trade-off` between recommended musical pieces. *Variety of artists* of the

[8] https://essentia.upf.edu/documentation/streaming_extractor_music.html.
[9] http://earsketch.gatech.edu.

recommended music is used as the measure to compute the diversity of recommendation lists. The solution proposed by the authors is a hybrid CF + CBF recommender system based on a *probabilistic generative model*, which contains the probability distribution over users, items, and features. It treats them as three separate conditionally independent probabilities. This model introduces a set of latent factors and weights associated with them. The content-based audio features used are *MFCCs*. The CF approach is based on a rating matrix using a three-point rating scale: disappreciation (0), neutral (1), and appreciation (2).

In Shao et al. [2009], the authors propose a recommendation strategy to solve the `accuracy-novelty trade-off` in recommending music items. They use *variety of artists* in music recommendation lists to define novelty. The authors argue that common similarity metrics used in music recommender systems do not capture the differences between the evolving and dynamic nature of music properties from (a) one type of music piece to another type or (b) a user with particular music preferences to another user. For instance, intensity is an important characteristic in measuring the similarity with rock music, while it is not so important with classical music [Shao et al. 2009]. Therefore, dynamic weights are specified to acoustic features based on genre. Also, the perception of the same pieces of music is often dissimilar for different users [Schedl et al. 2013]. Hence, it is also necessary to assign dynamic weights to audio features for different users to capture these subjective differences. The authors propose a *metric learning* approach to learn the best similarities according to the correlations between audio features and user consumption patterns with music items. The paper uses *hypergraphs* to merge consumption and content similarity data.

In Bu et al. [2010], the authors address the problem of `music recommendation in music communities` by *exploiting heterogeneous information* (social network information and acoustic properties of music). In typical online social communities related to music, such as Last.fm, each user can make friends with other users, listen to their favorite music tracks, team up to make playlists, join specific groups, and use keywords to bookmark music tracks, albums, and artists. The resources (music tracks, albums, and artists) can have relations with each other, e.g., a music track can be part of an album or a playlist. The approach proposed is named *music recommendation on hypergraph* (MRH), and its goal is to learn a *unified hypergraph* that can be used to both model multi-type entities and their relations in music social networks. A hypergraph is a generalization of an ordinary graph in which different relationships (e.g., music, tag, and users) are modeled via hyperedges, which are capable of capturing high-order relationships. The authors empirically explore the contributions of different types of social network information classified into four main dimensions: (1) social relations (i.e., friendship relations and group membership relations), (2) actions on resources (e.g., listening and tagging), (3) inclusion relations among resources (i.e., relations based on tracks and albums), and (4) acoustic-based music similarity relations for recommendation and show that MRH is helpful for practical music RS.

In Andjelkovic et al. [2019], the authors present a hybrid approach to `mood-aware music artist recommendation` that integrates artists' and users' mood as well as audio features. They propose a two-stage-recommender. First, candidate artists based on the target user's self-reported mood are identified, the latter being matched to expert-provided mood tags of artists, extracted from Rovi.[10] The resulting list of candidates is subsequently refined by a CBF component that leverages artist similarity based on audio content of the artists' most popular songs. This CBF layer performs reranking of candidate artists according to timbre, tempo, loudness, and key information.

In recent years, *deep neural networks* (DNNs) have been increasingly adopted for music recommendation, in particular in content-based systems to learn latent song or artist representations

[10] http://developer.rovicorp.com.

from the audio signal or from textual metadata (e.g., user-generated tags or artist biographies). Subsequently, these representations are commonly used in standard nearest neighbor CBF systems or as side information in CF-based approaches.

As one of the earliest works to adopt *DNN for music recommendation*, van den Oord et al. use a convolutional neural network (CNN) to represent each music item by 50 latent factors in the *latent user-item space*, learned from log-compressed Mel spectrograms of music audio [van den Oord et al. 2013]. They use the resulting latent factor representation of items together with latent user factors in a standard CF fashion and compare results to a weighted matrix factorization (WMF). The authors investigate two objective functions: mean squared error (MSE) and weighted prediction error (WPE). They found on a subset of the Million Song Dataset [Bertin-Mahieux et al. 2011] that CNNs substantially outperform linear regression trained on bag-of-audio-words created from MFCCs. Furthermore, using MSE as objective function outperforms WPE.

In Liang et al. [2015] the authors propose a hybrid architecture that uses probabilistic matrix factorization using audio content features as a prior to a probabilistic MF model, which is different from van den Oord et al. [2013], which uses content audio features to learn the *mapping* from audio features to user-item latent factors. They train a multi-layer perceptron (MLP) on vector-quantized MFCCs to *predict user-generated tags* (i.e., they define an auto-tagging task). The trained MLP is therefore assumed to capture semantics reflected in collaborative tags. The output of the last hidden layer is used as a latent content representation of songs, and the corresponding feature vector is used as prior to a Poisson MF. Evaluation experiments are conducted on a subset of the Million Song Dataset, both in WS and CS scenarios. In WS, the proposed approach performs equal to directly using the vector-quantized MFCCs. For a CS scenario, the proposed approach substantially outperforms the sole use of MFCCs.

A similar approach is followed by Wang and Wang [2014], who use a *deep belief network* (DBN). The authors extract spectrograms from the audio files under consideration, then apply principal components analysis (PCA), reducing feature vector dimensionality to 100, and finally feed the resulting feature vectors into a DBN. This allows the system to learn latent item representations, which are used as latent song factors in PMF. The key contribution of this work is to unify the two stages of feature extraction and recommendation in an automated process, using a model based on a DBN and a probabilistic graph-based model. Evaluation is carried out on a subset of the Million Song Dataset, which the authors enrich by song previews. Both CS and WS scenarios are considered and results are similar to those achieved in van den Oord et al. [2013].

In Oramas et al. [2017] the authors propose an approach that integrates latent song and artist representations to *create individual representations of music tracks and artists* and fuse both into a CBF system. The representations of songs are computed by feeding *vector-quantized spectrograms* into a *CNN*. Artist representations are learned using an *MLP*, which is fed with *TF-IDF* vector representations of artist biographies that had been enriched with information from DBpedia.[11] The resulting track and artist embeddings are then fused in a late fusion step, investigating two approaches: (1) concatenating the normalized embedding vectors and (2) using an MLP in which artist and track embeddings are connected to two different dense layers and fused to the output layer. The authors evaluate their approach for artist and for song recommendation on a subset of the Million Song Dataset. Results show that jointly considering the music recommendation problem at the song and the artist levels helps improve the quality of recommendations.

Music recommendation to enrich other media types: Aiming to `make user-generated videos more attractive`, the authors of Shah et al. [2014] propose a music recommender system

[11] https://www.dbpedia.org.

Table 1. Classification of *Audio Recommender Systems* According to Target Categories Sounds and Music

Target Audio		Paper	Item Content & Feature Extraction												Feature Postprocessing & Representation														Recommendation & Learning						
			basic signal-level audio features						feat. extr. method			meta-data		KG	text rep.	audio representation					fusion			semantic orientation			class			type				
			energy	spectral	timbre	melody	harmony	B, R, T	STFT	MFCC	CNN spect	editorial	user gen.		BoW/TF-IDF	Word2Vec	VQ	PCA	BoW	k-means	GMM	early	late	others	sig->semantic	sig->CF-LF	sig->NN	CBF	CF	CA	memory	LFM	graph	deep
Sounds	sound tracks	[Ostuni et al. 2015]										✓	✓	✓								✓						✓					✓	
		[Oramas et al. 2016]	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓								✓						✓	✓				✓	
		[Smith et al. 2019]		✓	✓				✓	✓																		✓	✓	✓*	✓			
Music	songs, artists, playlist	[Yoshii et al. 2008]		✓						✓										✓								✓	✓			✓		
		[Shao et al. 2009]	✓	✓	✓				✓	✓																		✓	✓				✓	
		[Bu et al. 2010]			✓					✓		✓	✓							✓								✓	✓				✓	
		[Andjelkovic et al. 2019]	✓	✓	✓		✓					✓	✓		✓									✓	✓	✓		✓		✓*	✓			
		[van den Oord et al. 2013]			✓					✓							✓	✓	✓							✓		✓	✓			✓		✓
		[Shah et al. 2014]		✓	✓				✓	✓			✓															✓		✓*				
		[Kaminskas et al. 2013]	✓	✓	✓	✓	✓	✓	✓			✓	✓	✓									✓		✓					✓				
		[Liang et al. 2015]			✓					✓			✓				✓								✓			✓	✓			✓		
		[Wang and Wang 2014]		✓					✓									✓										✓	✓					✓*
		[Oramas et al. 2017]		✓					✓		✓	✓	✓	✓	✓	✓	✓							✓			✓	✓							
		[McFee and Lanckriet 2011]	✓		✓		✓	✓					✓						✓	✓								✓					✓		
		[Vall et al. 2019]		✓					✓	✓	✓		✓			✓		✓	✓	✓	✓	✓			✓	✓		✓						✓

CA = √* refers to multimedia contextual factors; deep = √* refers to E2E; cf. Section 1.1.

that enriches user-generated videos (UGV) with a soundtrack that matches both video scenes and users' preferences. As a motivating example, the authors state that outdoor UGVs are often not appealing because of background noise. The proposed system leverages location, online listening histories, and past user activities and correlate them with the user's mood. Specifically, the *moods of UGVs are predicted* first based on the late fusion of geographical and visual features (color histogram) by relying on an approach named SVM-based probabilistic inference machine. Four mood classes are identified based on their level of *stress* and *energy* (low or high). Once the association between UGVs and scene moods is established, a list of songs is chosen as candidates for recommendation based on the predicted scene mood of the UGV. For this step, in an offline phase, the system fuses the visual and video features (*MFCCs*, *Mel magnitude spectrum*, and *pitch*) to evaluate the automatic selection of a matching soundtrack based on experience of professionals who create soundtracks for Hollywood movies. The user's listening history is used to further personalize the recommendation. The final result is an automatically generated music video that can enhance the user's experience because it contains music that matches scenes and locations.

Kaminskas et al. propose a context-sensitive music recommender for `location-aware music recommendation`, where locations correspond to points-of-interest (POI) [Kaminskas et al. 2013]. Locations are described by audiovisual material (images and descriptive text from Wikipedia), which constitute the media type to enrich in this case. The proposed hybrid approach integrates a knowledge-based component exploiting the DBpedia *knowledge graph* and an audio content-based approach to *music auto-tagging*. More precisely, given a POI as input, the knowledge-based recommendation engine estimates connection strengths between the target POI and the music pieces under consideration according to the nodes and edges in the DBpedia KG. The auto-tagging engine

uses *block-level features* (BLF) [Seyerlehner et al. 2010] that capture spectral, timbral, rhythmic, and tonal music characteristics to predict semantic labels (24 emotion words) for all music pieces in the collection. The matching between POIs and music items is then performed by computing the Jaccard index between the predicted emotion terms and human-generated annotations of the POIs using the same vocabulary. Both approaches (KB and auto-tagging) produce a score for all music items given a target POI. Results of the two recommendation engines are then fused by applying *Borda rank aggregation*. Evaluation shows that the combined approach outperforms its constituting components as well as a simple personalized baseline that recommends music of the user's preferred genre.

Automatic playlist generation: Another research direction in music recommender systems is automatic playlist continuation (APC). APC systems leverage sequential information reflected in users' listening sessions or playlists and use models created from this information to continue a given playlist. For more details on APC, please consider Bonnin and Jannach [2014], Quadrana et al. [2018], Zamani et al. [2018].

Exploiting item sequences in playlists can be achieved using *Markov chains*, as done in Chen et al. [2012], McFee and Lanckriet [2011]. While Chen et al.'s work does not use audio content features, McFee et al.'s work represents songs by timbre, loudness, tempo, and key descriptors extracted from audio, by tags, and by estimates of familiarity [McFee and Lanckriet 2011]. This information is used to train several Markov chains that model the transitions between songs in a collection of user-generated input playlists. The Markov chains are then used to generate recommendations for songs so they fit a certain playlist. The log-likelihood of the model for producing the actual playlists, as given in the ground truth, is used to measure performance.

Lately, *deep learning* is also used in APC. Many research works use *recurrent neural networks* (RNN) for sequence modeling in this case, however, not in combination with audio content features. Other variants of deep neural networks are used, for instance, by Vall et al. [2019], who propose profile-based and membership-based APC. The *profile-based* approach uses a *neural network classifier*, which categorizes each track with respect to its fit into a given playlist. The resulting candidate songs are then ranked according to their probability of matching the input playlist. The DNN-based classifier is capable of learning from any feature vector representation of tracks, and in turn derives the latent features. Among other variants of features, the authors consider the same *CNN* representation as used in van den Oord et al. [2013] as well as text embeddings created by *word2vec* [Mikolov et al. 2013b] from Last.fm collaborative tags. In addition to individual songs, the *membership-based* approach also describes playlists as feature vectors (of the included songs). In this approach, latent factors are derived not only for each song representation but also for the playlist features; the latter by averaging the DNN's latent song representations over all songs in a given playlist. As a result, songs and playlists are mapped to the same latent feature space, which enables computing the fit between a candidate song to add to a given playlist and the playlist itself by directly applying some distance metric.

Several recent approaches to APC use complex variants of *attentive neural networks*, e.g., Lin et al. [2018], Sachdeva et al. [2018]. Since they do not use audio content features, we refrain from discussing them here. For a detailed discussion of those, consider Schedl [2019].

2.2 Image-based Recommendation

In this section, we discuss approaches for image-based recommendation utilizing the categorization, whether visual content is used to recommend media products, e.g., images or paintings (Section 2.2.1), or non-media products that have an associated image, e.g., recommending clothing products based on the visual content of the product images (Section 2.2.2).

2.2.1 Image Recommendation. This section reviews literature that uses visual content extracted from images to recommend an image product. We classify this section based on the image type recommended according to: `generic photo recommendation`, `painting recommendation`, and `miscellaneous-image recommendation` (e.g., recommending dance background images and view recommendation while photographing).

Generic photos: Offering users support in finding a desired photograph from a collection shared or created by professionals or within a user community is an important task in a variety of domains and applications.

In Boutemedjet and Ziou [2006] the authors propose a hybrid filtering approach for image recommendation, which *combines a model-based CF and CBF* to address the `overspecialization problem` (attributed to CBF when the user is subject to recommendations that are highly similar to the consumed items by her) and the `new item problem` (attributed to CF for being unable to make recommendations when new items are added to the catalog that lack interactions/feedbacks hindering performance of CF). Hybridization can help to overcome the limitations in each individual model. The proposed system first represents images by RGB color histograms and classifies them into six classes using topic modelling based on the Dirichlet mixture model [Neal 2000]. A generative graphical model is used for CF using image content as side information, in which a probabilistic latent variable model combines users' preferences and items' classes into a unified framework. The authors propose to base the final image recommendation on a score computed from both predicted ratings and diversity of the recommendation list where diversity is computed according to differences in classes of visual documents.

Extending the previous work, in Boutemedjet and Ziou [2008] the authors propose an image recommendation model that improves their previous system by leveraging *context and metadata*. The proposed system is a `CA image recommendation system` that considers all the aforementioned factors in a unified recommendation model. In particular, a generative CF model is proposed that uses metadata and low-level visual information together with contextual information in which the role of visual content is to increase the diversity and novelty of recommended items. To provide recommendations, a utility is computed on the basis of the relevance of visual document to the target user in a given context expressed as a conditional probability. Results of a pilot study confirm the merits of the proposed system. In Boutemedjet et al. [2008], the authors demonstrate the significance of feature selection for image suggestion, in particular for rating prediction of high-dimensional image data.

Web `query recommendation` can be an important component for a user-oriented search engine. In recent years, recommending queries to search for trending images has emerged as an application of query recommendation [Westman and Oittinen 2006].

Trending image recommendation is composed of two phases: *trending event detection* and *trending image selection* [Yu et al. 2015]. However, most existing approaches are not tailored to the user's interests. To address this gap, Wu et al. [2014] propose a learning framework to provide `personalized trending image recommendation` according to users' preferences. The proposed system learns the user's interest from neighboring users (users with common searches) and then uses an MF model named trending-aware weight regularized matrix factorization (TA-WRMF) for the recommendation task. The system further incorporates *trending-aware visual features* including freshness and aesthetic quality of images. The evaluation is carried out on a large dataset of commercial search log (containing 21M users and 41M queries) and shows 50% gain in terms of query prediction accuracy with respect to baselines (CF-based and frequency-based baseline). The results largely support the usefulness of both personalization and trending-aware visual features for the problem of trending image query recommendation.

Further, given that media items are described by multiple modalities/features, an open research challenge is how to construct recommendation models for multimedia data in SM by merging the information obtained from different sources (e.g., textual and visual modalities) along with social interactions of users. To address this challenge, in Cui et al. [2010], the authors propose an approach for *fusing features of visual documents* and capturing correlations between features by introducing an effective feature representation structure named *feature interaction graph* (FIG). The proposed FIG is a two-level tree, where the root represents the target item (e.g., the image of an animal) and all the features are characterized by the leaf nodes. An edge exists between two nodes if there exists a correlation between them. Examples of item features include textual features (e.g., tags, titles, and comments), visual features (e.g., color, texture, and edge), and user features (e.g., the uploader of an image, the users tagging the image, or the users sharing it). The proposed FIG captures two types of correlations between features: the *intra-type* characterizing the correlation between features of the same type (edges between textual nodes, visual nodes, and user nodes) and the *inter-type* correlation characterizing the relation between heterogeneous features. The authors show the effectiveness of the proposed image RS by validating it on a real-life dataset from Flickr.

Motivated by the fact that conventional recommendation models, e.g., CF and CBF, face challenges such as extreme sparsity in image-sharing communities, in Niu et al. [2018], the authors propose an `image recommendation system for SM environments`. The proposed system is named *neural personalized ranking* (NPR), which extends BPR—a state-of-the-art rank optimization model for RS—by integrating it into layers of an NN and adding a nonlinearity layer to it.

The authors further propose an extension of NPR named *contextual NPR* (C-NPR) to integrate multiple categories of *side information* to overcome the sparsity issue. This side information includes: (1) geographical features, (2) topical features (information such as tags, title, and description of each image), and (3) visual features (obtained from CNN). Validation of the system on the Flickr YFCC100M dataset[12] shows the effectiveness of NPR, more specifically the contextual C-NPR model in comparison with several baselines including the BPR recommendation method.

It has been argued in some research works that recently—as a response to the users' emerging engagement in pure interest-based social services—the nature of social media has been shifted from user-centric social networks (SN), which are characterized by friendship or by following relationships, e.g., Facebook and Twitter, to *content-centric social curation networks* (SCN) such as Pinterest[13] and Delicious.[14] Users of SCN can explore and integrate interesting multimedia content into their "stories" for the purpose of creating and sharing experiences with other people and increasing consumption. SCN are characterized by the extreme *sparsity* of the user-image links and extreme *diversity* of the visual content. Due to these challenges, traditional RS may not be suitable for recommendation in SCN. In Geng et al. [2015] the authors propose a learning framework for `learning user and image features in SCN`. The proposed graph-based method exploits the connection between users and images and fine-tunes a CNN model to transform the heterogeneous representation of users and images into a homogeneous low-dimensional representation, thus effectively enabling CBF recommendation of images. For image representation, the system leverages ImageNet and AlexNet models [Krizhevsky et al. 2012]. The system is validated in extensive experiments on a large image dataset obtained from Pinterest with 1.4M images and 1M users.

A very recent research direction is *personality-aware recommender systems*, which consider the user's personality in the recommendation process [Schedl et al. 2018; Tkalcic and Chen 2015].

[12] cf. https://cacm.acm.org/magazines/2016/2/197425-yfcc100m/fulltext.
[13] https://www.pinterest.com.
[14] https://del.icio.us.

Personality is often modeled using the five factor model (FFM), which is based on the five personality traits of openness, conscientiousness, extraversion, agreeableness, and neuroticism (OCEAN) [John et al. 1991]. To build personality-aware RS, it is required to have access to information about users' personality traits. This information can be acquired through questionnaires, e.g., ten item personality instrument (TIPI) [Gosling et al. 2003] or big five inventory (BFI-44) [John and Srivastava 1999]. Alternatively, personality traits can be automatically inferred through supervised learning techniques, i.e., classification and regression, from user-generated data and individual cues [Azucar et al. 2018]. Such cues include choice of words in language, intonation of voice while speaking, and type of people one befriends or is in a relationship with [Guntuku et al. 2015a]. With images becoming a common medium for communication and expression, there is a recent trend to exploit users' favorite images shared on social media platforms [Ferwerda and Tkalcic 2018; Skowron et al. 2016] or even profile pictures [Celli et al. 2014] to recognize their personality. For a deeper discussion of personality recognition consider Azucar et al. [2018].

In Guntuku et al. [2015b] the authors study `the influence of users' personality` on their `preferred images`. They first extract a large list of features from a collection of user-liked images from Flickr (tagged as "favorite"). They propose to extend the traditional approach that identifies direct associations between the features extracted from an image and personality traits (features-to-personality) to a *two-step approach* in which the image features are first mapped to the answers of a *personality questionnaire (BFI-10)* and these answers are subsequently mapped to *personality dimensions* (features-to-answers plus answers-to-personality). The authors use a wide list of low-level visual features and consider adding features such as gender identification and scene recognition that might provide a better representation of the users' preferences [Cucurull et al. 2018]. The proposed personality-aware system is tested in an image recommendation system that offers users suggestions, which match their preference and personality profile. Results of the experimental evaluation show that using informative visual features and better personality modeling, it is possible to improve the personality estimation of users, which has a positive impact on the ultimate recommendation quality by the system.

Paintings: A number of research works have been carried out by Albanese et al. [Albanese et al. 2013] in the context of multimedia recommendation in on-line museums, where users can look to digital reproductions of paintings.

In Albanese et al. [2010, 2013] the authors propose a recommender system, which uses as input the observed painting and generates a list of recommended paintings as output. The recommendation model *combines consumption patterns* with *low-level visual features* (color, texture, and shape) and *image metadata* (painter, genre, subject, title) with the objective to distinguish users with comparative browsing behavior. The objective of their work is to define a collective user profile and a semantic representation of the items' content by modeling the recommendation as a social choice problem. The authors evaluate the system for recommending paintings in a virtual museum scenario (based on the Picasa dataset from the Uffizi gallery), and for recommending movies in a large dataset of movies. Experimental validation based on *effectiveness* and *usability* encourages further research in this direction.

In Bartolini et al. [2013] a recommender system is presented to provide `context-aware multimedia recommendations` in the domain of `cultural heritage`. Here, whenever a user is in the vicinity of a POI, the system provides personalized multimedia recommendation related to visiting paths for a given environment. The system uses *heterogeneous multimedia features* (low-level features, semantic metadata) and *contextual information*, and provides CA recommendations to give users access to multimedia services. A specific application of the proposed system was mentioned to be in the cultural heritage domain related to an outdoor scenario with a mobile user.

Table 2. Classification of *Image-based Recommender Systems* According to Target Type Image

Target Image		Paper	Item Content & Feature Extraction										Feature Post-processing & Representation													Recommendation & Learning						
			basic signal-level visual features				feat. extr. method				meta-data		text repres			image representation				fusion			semantic orientation			class			type			
			color	texture	edge	shape	SIFT/SURF	HOG	LBP	DWT/DFT	CNN	editorial	.	BoW/TF-IDF	Word2Vec	PCA	VQ	PCA	BoW	k-means	GMM	early	late	others	sig->semantic	sig->CF-LF	sig->NN	CBF	CF	CA	memory	model	graph	deep
Photos	trending search images SM images, generic photos	[Boutemedjet and Ziou 2006]	✓																								✓	✓			✓		
		[Boutemedjet and Ziou 2008]		✓	✓		✓					✓		✓					✓	✓		✓						✓	✓	✓		✓		
		[Wu et al. 2014]		✓			✓												✓									✓	✓	✓	✓		✓	
		[Cui et al. 2010]	✓	✓	✓		✓					✓	✓	✓					✓	✓				✓				✓	✓	✓*			✓	
		[Niu et al. 2018]									✓	✓	✓	✓		✓						✓					✓	✓	✓				✓*
		[Geng et al. 2015]									✓																✓	✓	✓				✓*
		[Guntuku et al. 2015b]	✓	✓	✓	✓				✓																		✓		✓				
Paintings	cultural heritage virtual museum	[Albanese et al. 2010]	✓	✓		✓						✓																✓	✓					
		[Albanese et al. 2013]	✓	✓	✓	✓						✓																✓	✓	✓*				
		[Bartolini et al. 2013]	✓	✓								✓	✓										✓					✓	✓	✓			✓	
Miscellaneous	photography, dance photo reminiscence	[Bourke et al. 2011]	✓																									✓		✓				
		[Rawat and Kankanhalli 2017]	✓	✓			✓	✓				✓	✓						✓	✓	✓	✓						✓	✓	✓				
		[Nguyen et al. 2016]					✓					✓	✓						✓	✓					✓			✓		✓*				
		[Wen et al. 2018]									✓		✓														✓	✓						
		[Peska and Trojanova 2017]									✓																	✓						

Note that the target product itself in all the reviewed research works in this table is an image. (CA = √* refers to multimedia contextual factors; deep = √* refers to E2E; cf. Section 1.1.)

Miscellaneous: In addition to the domains discussed above, there exist interesting niche domains in which image RS are used in a more general sense. We discuss a few in the following.

In Bourke et al. [2011] the authors propose a novel recommendation system for `mobile photography` with the aim to provide photography assistance to the user as she prepares to take photos. In particular, the authors build an Android app able to provide users with recommendations for well-framed photographs (by giving advice on position and framing), considering user's *current context* such as location, direction, and lighting conditions. The authors call the proposed system *social camera*. When the user points the camera on a scene, the social camera app provides a list of contextual features about the current scene not only limited to the current time and GPS coordinates but also compass direction, lighting conditions, as well as the current camera settings. The proposed system then provides users with recommendations of well-framed photographs. Over time, these suggestions help the user improve her photographic competence.

In a similar research line, [Rawat and Kankanhalli 2017] present a system for `viewpoint recommendation` with the goal to aid users in shooting high-quality photos nearby famous touristic locations. The proposed system is called ClickSmart and can provide real-time viewpoint suggestion to the user based on *the camera preview, time, weather, and location*. It leverages geo-tagged images and social media to learn the photo-taking behavior of other users. Quality of pictures is defined in terms of visual features such as RGB color histograms, HOG, and SURF features.

In Nguyen et al. [2016], the authors propose a photo RS for `supporting reminiscence for photo sharing on SM platforms`. Recommending photos from the past, similar to photos being shared in the present, will likely recall interesting and valuable memories. This phenomenon can be leveraged to enhance the user experience with the system. The proposed system is named NowAndThen and uses a variety of information sources such as *visual features* (based on Hessian-affine detector and SIFT descriptor) and *relations between users and tags* (people, locations, hashtags) to suggest photos most similar to user's current photos of interest, but potentially raising a more satisfying reminiscence.

In Wen et al. [2018], the authors propose an artistic work, a system for recommendation of visual backgrounds for dance performances that can help artists/dancers with the selection of images matching their dance style. The proposed system works by using deep matrix factorization (DMF) to combine the *dancer's actions on SM* with the *visual content* of the dance images shared by the dancer. The system extracts visual features from the dance-related images that the dancer shares on Pinterest. Then, it learns a user model based on the visual content of shared images to generate background visual recommendations.

2.2.2 Product Recommendation with Image-related Attributes. In our daily shopping experience, we make decisions that take into account how much the target products look *visually* appealing to us. Consider, as an example, the clothing domain. Most conventional RS are not (easily) usable for clothing products because clothing purchase has some special characteristics, that is people do not go with the crowd blindly when buying clothes nor do they buy the same clothing item twice [Sha et al. 2016]. Moreover, the aesthetic and visual characteristics of clothing are not necessarily reflected in commodity tags used in tag-based CBF systems. In the following, we report on RS that leverage the visual content associated with products for producing their recommendations.

Fashion: Humans naturally build up a sense of *relationships* between products that take into account their appearance [Prato et al. 2020]. Two types of relationships play a key role in our decision-making process: (1) *similarity* (or alternativeness), e.g., "Which pair of *jeans* matches these pair of *jeans* that I have liked?" and (2) *complementariness*, e.g., "Which *shirt* matches these *jeans*?"

To address both types of relationship, the authors of McAuley et al. [2015] propose a recommendation system that can recommend how to match clothes with accessories by exploiting the visual content extracted from images of these products. Their goal is to develop a visual and relational fashion recommender that is able to model the human sense of relationship between objects by utilizing the visual appearance of products. The proposed system solves a *network inference problem* defined on graphs of related images. Furthermore, instead of relying on hand-labeled images, the authors make use of freely available data collected from the Amazon web-store, containing over 180M relationships between a pool of almost 6M objects. Interactions identify pair of products that can be considered alternative and pair of products that can be considered complementary. The dataset, which the authors make publicly available, is categorized into top-level categories (books, clothing, movies, music, etc.) and finer-level categories (men, women, boys, girls, etc.). By incorporating the visual signals in the proposed system, the authors are able to discover various types of relationships between items beyond simple visual similarity and as such McAuley et al. [2015] can be seen as one of the first attempts to account for human preference in the appearance of a given object.

In He and McAuley [2016b] the authors propose a *factorization model* that combines visual features and users' preferences and apply it for recommending clothing and accessories using their associated images. The proposed method is named visual Bayesian personalized ranking (VBPR) from implicit feedback.

Characteristics of clothing that are considered as "fashionable" change as time progresses. For instance, the most fashionable women's sneakers change during each year/epoch. In He and McAuley [2016a] the authors extend VBPR to build a fashion recommender modelling fashionability over time by modelling both the *visual appearance* of products and their *evolution over time*. The proposed method is named temporally evolving visual Bayesian pairwise ranking (TVBPR) and exploits *deep CNN features* for modeling the *visual dimension* as well as the associated *temporal dynamics* (via incorporating information such as sales, promotions, or the emergence of new products). Integration of the non-visual temporal dynamic information into the

model allows to disentangle visual from non-visual decision factors and improve explainability of recommendations.

Product images of clothing provide a vast amount of information such as design, color schemes, decorative pattern, fabric thickness, and even quality of the product. Therefore, incorporating these cues in clothing RS can be a key to effective recommendation. Most research that acknowledges the importance of the visual appearance of clothing products on consumers' decisions either rely on pretrained CNN features or on standard visual descriptors, such as SIFT or color histograms, to represent images. CNN features, for instance, can be used to infer rich semantic information to classify products. As such, they can be useful to encode information such as "There is a skirt in the image." However, they are not designed to respond to the directive: "Find a clothing that is beautiful and matches the consumer's taste." The missing element here is `aesthetic understating of images` (see Datta et al. [2006] for a good introduction to the topic of aesthetic assessment of images). To bridge this gap, in Yu et al. [2018], the authors extract aesthetic information from images and incorporate it into a recommendation model. Specifically, the proposed system integrates an aesthetic network, i.e., a *brain-inspired deep network* (BDN) trained for image aesthetic assessment [Wang et al. 2016]. The BDN is used to extract holistic features of clothing products, for which it uses aesthetic ratings and photographic style labels for training. To acknowledge the varying taste of consumers in different time periods (e.g., different seasons), the proposed model utilizes a *tensor factorization model* for capturing diversity of aesthetic preference over time. Extensive experiments are performed on two real-world datasets: (1) Amazon.com [He and McAuley 2016b] containing clothing, shoes, and jewelry; (2) DPChallenge.com,[15] known as aesthetic visual analysis (AVA) dataset, which is the collection of images and meta-data introduced by Murray et al. [2012]. Results demonstrate the merits of the proposed aesthetic recommendation models with respect to non-aesthetic state-of-the-art recommendation models.

Mining the interpersonal `trust relationship` from users' social network can enable RS to solve the new user problem and create a new approach to enhance the performance of RS. In Sun et al. [2018], the authors propose a `personalized clothing recommendation` service that takes into account both *users' social circle* and *fashion style consistency* of clothing products. Fashion style consistency refers to the fact that two clothing items (e.g., tops and skirts) can be visually different but as long as they belong to the same style (e.g., sport, street, casual), the user may likely buy them; hence, fashion style consistency is an important element for the design of clothing RS. To this aim, motivated by previous research works [Qian et al. 2014], the authors of Sun et al. [2018] consider three factors in their proposed clothing RS: *interpersonal influence*, *personal interest*, and *interpersonal interest similarity*. In particular, five types of matrices are built by mining the social data available and other sources of information: S (representing user-user social influence), W (representing user-user similarity of interests), Q (representing user-clothing similarity), Y (representing clothing-clothing fashion style similarity), and R (representing user-clothing ratings). Afterwards, a probabilistic matrix factorization (PMF) framework is used to integrate the above observed matrices and recommend suitable clothing products to users by casting the problem in an optimization setting. The main role of the proposed PMF is to describe both users and clothing items in the same latent vector space, using the matrices Q, Y, R, S, and W. Evaluation is carried out on real-world datasets collected from Moguije,[16] a Chinese website for "social fashion" that blends a social network with online shopping possibilities.

An observation related to visual appearance of products is that people can have different levels of preference for different commercial items or parts thereof. In other words, different parts of

[15] http://www.DPChallenge.com.
[16] http://www.mogujie.com.

products may not contribute equally (or uniformly) to a user's preference when making a buying decision. To account for such inequality, in Chi et al. [2016] the authors propose a `visual part-based clothing recommendation`, which is able to decompose a given commercial item into a *set of disjoint partitions* where each partition characterizes a semantic component and *learns a part-based user model* (based on different partitions of the image) to obtain a personalized recommendation model. For example, a watch can be decomposed into two disjoint components: (1) the watch face and (2) the watch band. Each user can have different tastes regarding these two components when buying a watch. Experiments using the proposed recommendation model exploiting part-based visual appearance of items on a dataset from Amazon.com—including images of helmets, sports bottles, t-shirts, watches, and handbags—yielded improved results over existing textual or other visual RS that disregard appealing differences between parts of products.

In a similar work in Gu et al. [2016], `visual part-based image representation` is leveraged but to `model the user` herself. The proposed system accepts as input a frontal face photo and returns as output the best-fit eyeglasses. The proposed system does not rely on users' historical data but instead utilizes numerous facial attributes in an attempt to build the user profile directly from these `facial attributes`. Examples of such attributes are: *gender, race, eyebrow thickness, skin color, fatness, hair color* (17 fine-grained attributes), and additional ones that are learned through a set of *low-level visual features* via SVM and Adaboost. These facial attributes provide rich information about people and act as the main influence factor in the context of eyeglasses recommendation. Then, similar to facial attributes, seven frame attributes are defined, namely, type, shape, color, fit, materials, thickness, and size. The proposed recommendation framework relies on a *probabilistic graphical model* that leverages implicit matching rules between face and eyeglasses. Ranking of the frames (glasses) is realized based on pairwise similarity between user profile (the facial attributes) and the frames.

Other recently popular topics in the area of fashion RS include the use of graph-based learning approaches to learn user preferences at item-level or outfit-level [Li et al. 2020]; applications using attention mechanisms to estimate user preference, emphasizing on a fashion item in an outfit [Lin et al. 2020]; and ultimately explainable fashion RS, e.g., by building a fine-grained semantic space as presented in Hou et al. [2019].

In addition to clothing, multimedia recommender systems play a key role in recommending make-up products. In Chung [2014], the authors propose a `facial makeup style RS` to recommend makeup to women (e.g., eye shadow, blusher, eyelash, lipstick) according to their visual sensibility. The proposed system includes a *user interface, sensibility analysis, weather forecast,* and *CF* to satisfy the target user's needs in the cosmetic industry. In Alashkar et al. [2017a], a rule-based facial makeup recommendation service is proposed that is able to provide makeup style recommendations based on *occasions, trends,* and the *user's facial features*. The system is based on a KB that correlates facial features with makeup styles. In Alashkar et al. [2017b], the authors integrate domain knowledge represented as examples-rules with a *deep neural network*–based makeup recommendation model. To this end, the cost function of the deep neural network is modified to fit both human-annotated and rule-annotated labels. Human facial traits are fed into the network, which leverages the pairwise *Before-After* images and the makeup artist knowledge jointly. The authors demonstrate the capacity of their system to recommend homogeneous makeup styles that match human facial traits.

Food: A considerable portion of searches on the web are related to food or lead to food-related websites. Providing people access to online recipes or making personalized food recommendations can assist people in selecting dishes that suit their palate and are healthier. Food RS are typically designed for one of the following purposes: (1) to model users' food preferences and present

individuals with meal recommendations and (2) to assist users in consuming healthier food [Trattner et al. 2018]. However, making recommendations that satisfy both food preferences and nutritional requirements is generally a challenging task [Yang et al. 2017]. Recent findings in the psychology of human decision-making suggest that the visual nature of food choice (e.g., the uploaded images of recipes) has a high impact on users' preferences and choices. In the following, we review research works that consider visual signals for development of food RS.

In Elsweiler et al. [2017] the authors present a study with the goal to understand whether or not users will select `healthy foods suggested by a food recommender`. The authors use a multimodal recipe dataset composed of recipe names, images, ingredients, nutritional information, and recipe popularity. The authors use the dataset to investigate how people perceive and select recipes. Results indicate that it is possible to predict the preferred recipes of users with good performance if low-level image features and recipe meta-data are used as features. As an interesting finding, the study further reveals that many of the recipe choices made by users are driven by the visual effect of images.

Acquiring user feedback about food using traditional interfaces is difficult. To address this problem, in Yang et al. [2017] the authors `propose a visual interface` composed of images for preference elicitation in the food domain. The proposed model is called *FoodDist* and can learn a *food image embedding* using CNNs. It uses dietary profiles and *multitask learning* to personalize meal recommendations. FoodDist uses Euclidean distance embeddings of food images, which are shown to be capable of `finding the most similar foods` for a given food image. Interested readers are referred to Herranz et al. [2018] for a review on how visual content and other information such as context and external knowledge can be integrated into food-oriented applications such as recipe analysis and food recommendation.

Tourism: With the significant increase of smart phones and location-based social networks (LBSN) such as Foursquare,[17] Gowalla,[18] and Brightkite,[19] a large number of users can easily share their experience with points of interests (POI) via the so-called "check-ins." The availability of users' check-in data has created new opportunities for designing assisting services that can facilitate users' travels and social interactions. One such service is POI recommendation service (a.k.a location recommendation), which aims to `recommend new POIs` to users according to their personal preferences to facilitate their exploration of new areas of a city and also helps advertisers to `provide location-based recommendations`. The majority of related works on POI recommendation consider the following four factors as the ones effective for improving recommendation quality: (1) *geographical influence*, i.e., users intuitively tend to visit nearby POIs; (2) *social influence/correlations*, i.e., the choices of friends may contribute to the decision-making; (3) *temporal patterns*, i.e., users' check-in preferences depend on the hour of the day; and (4) *textual content indications* such as user-generated tags, sentiment, or POI properties. In modern LBSN, users can post photos associated with their locations. Photos contain a rich source of information that reflect users' interests. For instance, a user posting photos of monuments is likely to visit historical locations or museums, while a user who posts images about water sports could be more interested in visiting seaside locations. Therefore, images can improve the quality of POI recommendations. This section focuses on application of visual content for RS in the tourism domain.

In Wang et al. [2017] the authors investigate the `impact of visual content` of the photos shared by users in modelling their taste for `POI recommendation`. They propose a new recommendation framework named *visual content enhanced POI recommendation* (VPOI). VPOI uses *CNNs*

[17] https://www.foursquare.com.
[18] As Gowalla was online until 2012, we provide the wiki page to it: https://en.wikipedia.org/wiki/Gowalla.
[19] https://www.brightkite.com.

Fig. 3. Examples of indoor scenes/food in an American restaurant (Figures (a) and (b)) and Japanese restaurants (Figures (c) and (d)), showing different styles of restaurants in terms of visual appearance in scene and food. For example, American style restaurants are bright, colorful, and warm; while Japanese-style restaurants are neat, simple, and clean. Furthermore, American dishes use large amounts of food or lots of fries, while Japanese food has a refined and elegant plate presentation. (Picture taken with permission from Chu and Tsai [2017].)

pretrained on ImageNet to extract visual features from images and uses *probabilistic matrix factorization* (PMF) to model the interactions between visual content, users, and locations. Experimental results on real-world datasets of images from New York City and Chicago shared on Instagram between October 2015 and February 2016 show the usefulness of the proposed system particularly in coping with CS by incorporating images. The proposed CNN-based approach is also shown to outperform the use of SIFT and HOG descriptors.

In Chu and Tsai [2017], the authors study the impact of `food/location images in restaurant recommendation`. The contributions of this research work are: (1) to seamlessly and finely integrate visual information into the restaurant recommendation framework and (2) to build a hybrid recommendation system aided by visual information. For (1), the authors use photos in blog articles in addition to textual information to model restaurant attributes and user preferences (see Figure 3 for illustration). Images are classified into *four categories* (indoor, outdoor, food, and drink) using a pretrained CNN and a SVM. In addition, colors are extracted from images in each category. As for (2), *hybrid recommendation* models are built by enhancing MF and BPR-MF [Hu et al. 2008] and leveraging the visual content. Experimental results on a dataset from a restaurant-dedicated social platform in Taiwan[20] show that the CNN outperforms baselines such as the one based on textual features or on visual information using color names. In addition, pre-classification of images improves recommendation quality.

A related research line that has attracted significant interest in tourism and advertisement industries is the problem of `venue category recommendation`. Its goal is to suggest appealing localities within close proximity to users' current location. To address this problem, in Farseev et al. [2017] the authors propose a recommendation framework across different data sources (a.k.a cross-domain recommendation), which exploits multiview SM data. To this end, Foursquare venues are recommended to users with accounts also on Twitter and Instagram. By leveraging different *venue category preferences* with location indicators, the system recommends a good-fitting venue

[20] http://hungry.9ifriend.com/main.

Table 3. Classification of *Product Recommender Systems Leveraging Visual Content* According to the Product/domain

Target Product		Paper	Item Content & Feature Extraction											Feature Postprocessing & Representation														Recommendation & Learning						
			basic signal-level visual features				feat. extr. method					meta-data		text repres	image representation						fusion			semantic orientation			class			type			
			color	texture	edge	shape	SIFT/SURF	HOG	LBP	DWT/DFT	CNN	editorial	user gen.	BoW/TF-IDF	WordzVec	PCA	VQ	PCA	BoW	k-means	GMM	early	late	others	sig->semantic	sig->CF-LF	sig->NN	CBF	CF	CA	memory	model	graph	deep
Fashion	clothing, make-up, style, eyeglasses	[McAuley et al. 2015]									✓															✓	✓		✓*			✓	
		[He and McAuley 2016b]									✓															✓	✓	✓				✓		
		[He and McAuley 2016a]									✓															✓	✓	✓				✓	✓	
		[Yu et al. 2018]	✓								✓	✓										✓				✓	✓	✓	✓*		✓			
		[Sun et al. 2018]									✓	✓														✓	✓	✓			✓			
		[Chi et al. 2016]	✓	✓		✓	✓												✓	✓		✓						✓		✓*				
		[Gu et al. 2016]	✓			✓		✓	✓	✓		✓																✓		✓*				
		[Hou et al. 2019]									✓	✓	✓									✓			✓			✓	✓					✓
		[Li et al. 2020]									✓	✓													✓		✓	✓	✓				✓	✓
		[Lin et al. 2020]									✓	✓												✓	✓		✓	✓	✓	✓*			✓	✓*
		[Chung 2014]	✓	✓								✓																✓	✓	✓*				
		[Alashkar et al. 2017a]	✓			✓		✓	✓			✓										✓						✓		✓*				
		[Alashkar et al. 2017b]	✓	✓		✓		✓	✓			✓										✓						✓		✓*				✓
Food	recipe, dish	[Elsweiler et al. 2017]	✓	✓								✓	✓	✓								✓						✓						
		[Yang et al. 2017]									✓	✓	✓												✓		✓	✓						
Tourism	hotel, POI, restaurant	[Wang et al. 2017]		✓	✓		✓	✓			✓																✓	✓	✓	✓		✓		
		[Chu and Tsai 2017]	✓										✓	✓								✓			✓		✓	✓	✓	✓		✓		
		[Farseev et al. 2015]									✓	✓	✓	✓										✓			✓	✓	✓				✓	
Info	tags	[Lin et al. 2011]	✓	✓	✓		✓			✓			✓	✓														✓	✓	✓*		✓		
		[Liu et al. 2014]					✓						✓	✓					✓					✓				✓		✓*		✓		

Note that the targets here can be anything from commercial products (e.g., clothing) to food or some form of information given to the user (e.g., on points of interest). (CA = √* refers to multimedia contextual factors; deep = √* refers to E2E; cf. Section 1.1.)

near the user's current location, instead of better-fitting venues located far away from the user. The proposed framework utilizes multi-source multi-view data such as user-generated content extracted from texts and images, and via location processing techniques. Experimental validation of the framework is carried out on a large multi-source multi-modal cross-region social dataset named NUS-MS [Farseev et al. 2015] and shows the merits of the system over state-of-the-art baselines.

Tag/Annotation: Tag recommendation is a process that facilitates the creation of free-form text descriptors for music, pictures, and so forth, usually by selecting preferred keywords from a set of candidates. Tags make it easier for users to search for and to organize multimedia contents. Two types of tag recommendation systems can be distinguished: (1) generic tag recommender systems (a.k.a automatic annotation systems) and (2) personalized tag recommendation systems.

In Lin et al. [2011], an `automatic image annotation system` is proposed that transforms the task into a personal item recommendation problem by considering *images as items* and *users as tags*. Then a rating prediction problem is solved via a standard TrustWalker model, which aims to combine trust-based and item-based recommendation. In particular, the algorithm attempts to search the trust set of the target user (i.e., users in the target user's trust set) and return their rating directly. If users have not rated the target item, the algorithm returns ratings of similar items based on their visual similarity or further extends the search. For calculating visual distance between images, six types of features for each image are extracted: color layout, Gabor filters,

SIFT, scalable color, among others. Preliminary validation of the system is carried out on two image datasets from the University of Washington[21] and the MIRFlickr dataset [Huiskes and Lew 2008].

In Liu et al. [2014], the authors build a `personalized tag recommender for photos shared in SM`, which unifies *user preferences and geo-location* to define the *most relevant tags*. The authors assume that different users and different geo-locations have different preferred tags associated with an image. To represent the visual content of photos, bag-of-visual-words using SIFT descriptors are used. Given a user's untagged photo and its geo-location, the model predicts both user-preferred and geo-location-specific tags. Then, the visual appearance of the photo and the predicted tags are combined to discover related photos, whose most frequent tags are recommended. Results show that learning user preference and geo-specific preference are important tasks for accurate tag recommendation.

2.3 Video Recommendation

Video recordings are complex signals composed of different modalities, in particular, the audio and visual modalities. The impression of a video by a user is impacted by numerous elements related not exclusively to its content yet additionally to the general video style, reflected in its sound and visual content. For instance, the two movies *Schindler's List* and *Empire of the Sun* are from the same genre and director—both are emotional movies coordinated by Steven Spielberg and portray historical events—but their styles vary substantially, the former shot documentary-like in black-and-white and the latter making use of special effects and bright colors. These distinctive characteristics of multimedia meet users' different information needs [Deldjoo et al. 2019; Neumayer and Rauber 2007] (e.g., when watching a documentary) or entertainment needs [Schedl et al. 2013] (e.g., when watching an action movie).

In this survey, we classify research on video recommendation into works that target *movie recommendation* (Section 2.3.1) and works that target *user-generated video recommendation* (Section 2.3.2), where the latter is performed mostly in the context of social media.

2.3.1 Movie Recommendation. In the following, we review papers that recommend the user a movie by accounting the audio-visual characteristics of the underlying content. One of the aspects of almost all reviewed works is that they use a media *derived from movies*, such as trailers, movie clips, or even posters or a few selective frames, for content analysis in a movie recommendation system. The reasons are manifold: (1) feature extraction/content analysis from full-length movies is computationally much more demanding than from short trailers; (2) for user studies, it is time-consuming to ask a user to watch the full-length-movies; and (3) full movies are not (always) freely accessible. These issues can be properly addressed by using a representative trailer version of the movies as performed in the following works; however, whether features extracted from movie trailers or posters are representative of full movies remains an open question that needs to be studied in more depth.

The user interested in a movie X may like movie Y because they have similar posters and similar frame scenes. In Chen et al. [2018], the authors propose a movie recommendation system that leverages visual content information in movie posters and selected movie frames to enhance recommendation quality. The proposed system is an *E2E architecture* (cf. Section 1.1), which unifies the two problems of visual feature extraction (from a pre-trained CNN) and recommendation into a unified optimization process, named unified visual contents matrix factorization (UVMF).

[21] http://www.cs.washington.edu/research/imagedatabase.

In Deldjoo et al. [2016b] the authors propose a CBF movie recommendation system that relies on a set of `stylistic visual features` defined as *shot length*, *color variation*, *lighting key*, and *motion vectors* to filter movies. For instance, comedy movies are usually made with a large variety of bright colors while horror films use dark hues; this property can be captured by the visual feature of color variance. Several improvements of the proposed movie RS were presented by the authors in Deldjoo et al. [2018a, 2016a, 2018c], among others, experiments with a more diverse set of visual features—with a much higher dimensionality—such as the ones based on MPEG-7, pre-trained CNNs, and aesthetic visual features.

We would like to point to a recent thorough study [Du et al. 2020], in which the authors test the impact of a comprehensive set of content features (audio, visual, and textual) to build a movie RS. One key insight of their approach, named *collaborative embedding regression* model, is that whenever one of the content features is not available, the model can combine arbitrary other features with interaction data to provide effective recommendation. The model also uses a *priority-aware late fusion* method to boost the quality of movie recommendations.

In the following, we review a number of research works that actually validate the functionality of their systems for recommendation in *CS scenarios*.

In Zhu et al. [2013], the authors propose a CBF movie recommendation framework named VideoTopic, which breaks down the recommendation problem into *topic-based item representation and recommendation*. In the first stage, the proposed method exploits visual and textual features of movies and constructs a topic model based on LDA to represent item contents as well as the user interests, where the latter is computed from the topics of users' recently watched videos. For the textual features, two types of movie metadata, actors and directors, are used claiming that among other movie metadata (plot, actors, director, writer), the former contain richer information [Gantner et al. 2010]. For the visual information, the system extracts SIFT visual features from trailers. Visual features are extracted from three key frames of each trailer. In the second stage, the recommendation is formulated as finding items with minimal topic distribution difference with respect to users' interests. The benefit of the proposed systems is that the interests of new users are learned on the fly as they watch movies, while for the existing users, their current interest is measured based on their old interest and the current movie being watched. The evaluation on the MovieLens 1M dataset shows that for new-item problem, VideoTopic outperforms other methods using unimodal textual or visual features with respect to accuracy metrics.

In Roy and Guntuku [2016], the authors present an `emotion-based visual movie recommender system` named visual-CLiMF (Collaborative Less-is-More Filtering). Visual-CLiMF extends the original CLiMF approach [Shi et al. 2012] and further combines visual factors with consumption information to learn a latent representation that relates to emotive factors. Experiments on a video dataset made by the authors and named video emotion dataset (contains 323 movie/TV series clips)[22] and the Amazon products dataset containing product reviews and product metadata (CNN-based image features) demonstrate that visual-CLiMF can outperform existing CF methods or the ones without (CLiMF) content information.

Other works integrate *psychological aspects* when making recommendations, most commonly personality traits or affective cues that are extracted from metadata about items, e.g., reviews or comments. This seems highly reasonable, since such aspects are known to influence human preferences, e.g., for music [Ferwerda et al. 2017], movies [Golbeck and Norris 2013], or books [Rentfrow et al. 2011]. While we are not aware of any work that joins personality information and movie content in a CBF, *affect* is considered, for instance, in Benini et al. [2011], Canini et al. [2013]. In Canini et al. [2013] an *affective framework* is proposed for extracting audio-visual features (e.g., dominant

[22]Dataset will be made available by authors upon request.

color, color energy, lighting key, sound energy, low-energy ratio, MFCCs) and *movie grammar descriptors* (e.g., illuminant color, shot length, and shot type transition rate) allowing movie scenes to be compared based on their emotional difference, which are recommended eventually.[23] The proposed model allows linking between video content features to users' emotional preference. Interested readers on affective multimedia system can refer to the survey Wang and Ji [2015] and to Guntuku et al. [2016] for further information on impact of personality and culture background on users' affective response to multimedia content.

In Deldjoo et al. [2019], the authors describe a movie recommender system for the new item cold-start problem. The system makes use of canonical correlation analysis to combine *audio and visual features together with movie metadata (genre and cast)* into a unique representation (the *Movie Genome*). The proposed system also exploits a hybrid recommendation model named collaborative-filtering enriched CBF (CFeCBF) on its core, which is designed to train a CF model on warm items and leverages it on the movie genome to recommend cold items. Results on a dataset of trailers named MMTF-14K [Deldjoo et al. 2018b] indicate that multimodal recommendations generated by the proposed CFeCBF model significantly outperforms pure CBF baseline using genre or cast as item contents.

2.3.2 User-generated Video Recommendation. As one of the earliest approaches in video recommendation, Mei et al. [2007] and Yang et al. [2007] introduced VideoReach, an online video recommender system that uses a *multimodal relevance* assessment of new videos with respect to a selected video in addition to users' click-through data to compute recommendations. The authors use an attention *fusion function (AFF)* to combine the relevance scores from different modalities. Furthermore, since various videos may have different importance weights related to the three modalities, *click-through data* are exploited to automatically adjust weights for modalities and videos. The most recent version of VideoReach can be seen in Mei et al. [2011], which moves the previous versions forward by: (1) formulating the multimodal relevance as finding a suitable set of pairs, constituted by a set of functions (e.g., L_1 distance and AFF) and corresponding weight; (2) exploiting users' watching behaviors (e.g., watching a specific video segment) to estimate the corresponding feature weight; (3) performing visual concept detection to estimate the video category; and (4) evaluating effectiveness in comprehensive experiments. One of the key assumptions of these works is that textual information has more importance compared with visual and audio information; hence, videos with low textual relevance are filtered out by the AFF. VideoReach leverages different types of user interaction patterns; for example, what is the next video the user clicks, whether this user played, skipped, or browsed fast. Evaluations over a large dataset of online videos from MSN[24] show the usefulness of the proposed system as an alternative to CF models in CS scenarios where user interactions are unavailable or insufficient. A similar multimodal framework is proposed in Luo et al. [2009] for *news video* recommendations.

With the advance of *social networks* (SN), the number of research works that have leveraged *social data* to improve the performance of *video recommendation systems* has been increasing. In Zhao et al. [2012a], the authors propose a social video recommendation system that calculates a score for each candidate video with respect to a reference video, where this score consists of two components: (1) how much the user's friends are interested in the video and (2) the similarity between the user and friends. The *friend-content utility* is calculated based on *textual, visual,* and *popularity scores*, which are linearly combined (the weights set empirically). As for the popularity, different definitions are considered, based on "total views," "favorites," "rating," and "comment

[23] Some of the features are described using mean and standard deviation of features over frames.
[24] https://www.msn.com/en-us/video.

count." The *friendship preference similarity utility* between users is calculated by measuring the distance between the *commonality between tag sets* of users (the tags of videos viewed by two users). The final recommendation score combines the above by using a product aggregation function. The authors of Zhao et al. [2012b] present a similar work to Zhao et al. [2012a] but instead replace friend similarity estimate with *relationship strength estimate* between the target user and their friends in different domains using a graph-based approach.

The authors of Cui et al. [2014] propose regularized dual factor regression (REDAR), a matrix factorization framework that `leverages social attributes of users` to make recommendations. The proposed system addresses the limitations of traditional latent factor models (i.e., latent factors are hardly interpretable and the model suffering CS settings for both new user and item). It represents users by their *usage of content attributes of videos* (textual description and visual information) and *videos by users' social attributes* (demographic information and produced tags). Thereby, both videos and users can be represented in a shared space in which latent factors are influenced by social attributes and content attributes. The usefulness of the proposed system is validated via extensive experiments in a real SN dataset named Tencent Weibo (a Twitter-style social network platform in China); the results indicate that in a majority of cases the proposed method outperforms existing baselines with relative improvement of 20% or more.

`Cross-network user modeling`, which focuses on aggregating users' information from different platforms, can create recommendation opportunities not only for enriching the existing recommendation quality by better understanding users' interests but also to address existing RS issues such as data sparsity and CS. The latter can be solved by integrating users' behavioral patterns among OSN. For example, if one can identify new users on other well-established OSN sites, we can transfer knowledge from the mature OSN to the new OSN platform, thereby alleviating data sparsity or new user CS problem. In Deng et al. [2013], the authors propose a `YouTube video recommendation solution` by incorporating user information from Google+. The proposed system is composed of two main steps: (1) *profile enrichment*, whose goal is to enrich user profiles from the information coming from auxiliary platforms (Google+); and (2) *collaborative relationship transfer*, whose goal is to model users' similarity in terms of their behavior. The considered behavioral aspects include users' active actions (e.g., "upload," "favor," or "adding to a playlist") on videos and content information associated with videos (e.g., tags, video categories, and visual features represented via SIFT descriptors).

In Ma et al. [2018], the authors present LGA, a latent genre-aware `micro-video recommendation model based on SM information`. Micro-videos have a duration between 6 and 300 seconds and are created by users of OSN. The peculiarity of a micro-video is to convey in this extremely short time a self-contained and clear thought, image, or idea.[25] LGA extracts *interaction features* as well as *auxiliary features* describing context and visual content. These features are later fed into a *deep-learning model* used to learn both the latent genres of items and the recommendation scores. To validate their approach, the authors created a real-world micro-video dataset collected via the Twitter Streaming API. They show the merits of the system with regards to both effectiveness and efficiency measures.

Some works leverage information about the users' *emotions* to build an `emotion-aware short film recommendation`. For example, the authors of Orellana-Rodriguez et al. [2015] consider user comments of YouTube video clips that represent short films. From these comments, emotions are extracted using a lexicon-based approach. The authors consider four contrasting pairs of emotions: joy–sadness, anger–fear, trust–disgust, and anticipation–surprise. The identified emotion terms are weighted to create an emotion vector per video clip, which can be regarded as a content

[25] https://www.techsmith.com/blog/introduction-to-microvideo.

Table 4. Classification of *Video Recommender Systems* According to Target Type Video

Trg. Vid.	Paper	Item Content & Feature Extraction									Feature Post-processing & Representation												Recommendation & Learning						
		audio feat. extr. method			visual feat. extr. method				motion	meta-data		text repr.			image and audio representation					fusion			semantic orientation			class			type		
		basic signal	STFT	MFCC	basic signal	SURF/SIFT	MPEG-7	CNN		editorial	user gen.	BoW/TF-IdF	Word2Vec	PCA	VQ	PCA	BoW	k-means	GMM	early	late	others	sig->semantic	sig->CP-LF	sig->NN	CBF	CF	CA	memory	model	graph	deep
Movie	[Chen et al. 2018]							✓																✓		✓	✓				✓*
	[Deldjoo et al. 2016b]				✓				✓																	✓			✓			
	[Deldjoo et al. 2016a]				✓				✓																	✓	✓			✓		
	[Deldjoo et al. 2018]						✓	✓												✓					✓	✓	✓			✓		
	[Du et al. 2020]			✓		✓		✓	✓	✓	✓	✓				✓	✓				✓				✓	✓	✓			✓		
	[Zhu et al. 2013]					✓				✓		✓					✓					✓				✓		✓*		✓		
	[Roy and Guntuku 2016]							✓		✓						✓							✓			✓	✓	✓		✓		
	[Canini et al. 2013]	✓		✓	✓		✓		✓	✓												✓				✓			✓		
	[Deldjoo et al. 2018a]		✓	✓	✓			✓		✓		✓							✓		✓					✓			✓		
	[Deldjoo et al. 2019]		✓	✓	✓			✓		✓		✓							✓	✓						✓	✓			✓		
UG Video	[Mei et al. 2011, 2007; Yang et al. 2007]	✓			✓				✓	✓	✓	✓								✓		✓				✓		✓*	✓		
	[Zhao et al. 2012a,b]				✓					✓	✓	✓										✓				✓	✓		✓			
	[Cui et al. 2014]					✓				✓	✓	✓										✓				✓	✓			✓		
	[Deng et al. 2013]					✓				✓	✓	✓					✓	✓		✓	✓					✓	✓				✓	
	[Ma et al. 2018]							✓		✓		✓								✓		✓			✓	✓	✓					✓
	[Orellana-Rodriguez et al. 2015]										✓	✓											✓			✓	✓	✓	✓			

CA = $\surd^*$ refers to multimedia contextual factors; deep = $\surd^*$ refers to E2E; cf. Section 1.1.

feature describing the item (clip). For recommendation, the authors employ a collaborative ranking approach realized by learning-to-rank using LambdaMART. The authors evaluate the similarity of the emotional context automatically associated with a short film with those explicitly annotated by humans on a dataset collected from two popular short film festivals available on YouTube. Results indicate similarities between two different emotional context approaches (automatic vs. manual); however, the automatic approach is capable of predicting an affective context for emotion-aware personalized ranking.

3 SUMMARY AND FUTURE DIRECTIONS

Leveraging multimedia content in recommender systems is a highly relevant problem in practice and it is not limited to domains in which recommended items are media items: multimedia content has been widely used in the past five years to recommend non-media products, mostly in the context of e-commerce or fashion domains.

In this survey, we reviewed and categorized the RS able to leverage multimedia content developed over the past 15 years. In particular, in this literature review, we have introduced a taxonomy of multimedia recommender systems that categorizes them according to the main domains and tasks in which RS benefit from multimedia content analysis. Instead of focusing on one particular media type (e.g., images), this work is the first comprehensive survey that organizes, analyzes, and comments on the recent literature according to the most common media types targeted in RS research: audio, visual (i.e., images), and video. Furthermore, we categorized the reviewed research works from the the perspective of the core recommendation technique, usage of different modalities, content-based features, and types of metadata, among others.

3.1 Lessons Learned

Among different lessons learned by this literature review, we highlight a few prominent ones below:

Feature extraction. RS leveraging multimedia content are still dominated by handcrafted approaches, where (1) features relevant to the recommendation task are identified by domain experts and (2) the recommender algorithm is based on rule-based approaches. Overall, less than one-third of the reviewed works use deep learning features. Even though higher human effort is required to design a recommender system with handcrafted features, the advantage over deep learning approaches is that recommendations generated with handcrafted features or ad hoc algorithms are easier to explain and, as such, increase trust in the system [Pu and Chen 2007]. Moreover, indications exist in other application areas that the progress of deep learning models over ad hoc handcrafted approaches is not always as substantial as expected [Dacrema et al. 2019; Lin 2019].

The main scenario in which the usage of deep learning features and algorithms is predominant is that of *product recommendations leveraging visual content*, a scenario that greatly overlaps with that of traditional e-commerce recommender systems. The predominance of deep learning approaches in this scenario is partially motivated by the abundance of information produced by users on items, in the form of interaction logs. Deep neural networks are flexible in learning useful representations whenever a large amount of information about items and users is available and the design of ad hoc models and handcrafted features become infeasible [Zhang et al. 2019]. The advantages of using deep learning in this scenario are two-fold: (1) it enables automatic feature learning from a huge amount of raw data, and (2) it enables recommendation models to include both multimedia features and user-item interactions.

The application scenario of product recommendations leveraging visual content is dominated by *fashion* use cases. Fashion, by definition, is a style that is popular at a particular time among a category of users. In the fashion domain, the user may not buy an item uniquely for its content similarity with another item, but also because it is compatible with other fashion items within an outfit, e.g., blue pair of pants compatible with a white shirt. A proper evaluation of this type of compatibility (a whole outfit whose items share some stylistic properties) lies outside the capabilities of handcrafted features and requires a deep integration of visual features and user behavioral analysis, an integration empowered by deep learning techniques.

Semantic orientation. From a pure machine learning perspective, many approaches have been designed able to directly predict the relevance scores of given user, item pairs to produce personalized recommendations. Although these methods have the potential to lead to higher offline accuracy, from an industrial and application-oriented perspective, the most successful approaches are those that introduce semantic-oriented and in particular human-interpretable tagging as an intermediate step between the extraction of multimedia features and the generation of recommendations. Turning multimedia features into structured tags is important for both editors and curators, as structuring the content supports users' information consumption and allows for a better tuning of the recommender system.

Beyond-accuracy. Traditional content-based recommender systems tend to frequently recommend the same set of items to the users and result in similar recommendation lists and low diversity, whereas a recommender better able to diversify its recommendations will exhibit higher utility for the users [Zhou et al. 2010]. In this regard, multimodal recommender systems, as well as systems leveraging visual or aural modalities, improve beyond-accuracy performance of recommendations and avoid the so-called filter bubble issue [Haim et al. 2018].

3.2 Open Research Questions

We have identified several open research questions, which we put forward for the further evolution of MMRS:

Multimodal recommender systems. Given that media items are described by multiple modalities, an open research challenge becomes how to design recommendation techniques for multimedia

data that are able to properly combine different sources of information (e.g., textual, audio, and visual). Learning from multimodal sources offers the possibility of uncovering relationships between modalities and gaining an in-depth understanding of natural phenomena [Baltrusaitis et al. 2019].

Leveraging additional interaction types. Generally, in the context of recommender systems leveraging multimedia content, researchers rely on a few types of user-interaction types with media items, such as viewing a video or listening to a soundtrack. In real-world applications, however, there are many additional actions that can be performed on certain media items (e.g., zoom in on a specific area of an image, re-watch a scene in a video, increase the volume when listening to a part of a song) that are not considered in today's research.

Security and privacy of multimedia recommender systems. A recently popular task is to study the impact of machine-learned adversarial attacks against latent factor models. Several recent works have reported the vulnerability of MMRS against such attacks, e.g., Di Noia et al. [2020], Tang et al. [2019]. Compared with handcrafted shilling attacks [Deldjoo et al. 2020b], in adversarial attacks, the goal of the attacker is to design norm-constrained adversarial perturbations to be added to the data (embedding or raw content) to alter recommendation results toward an engineered outcome. Studying the characteristics of these modern attacks is vital to design effective countermeasures against them. For further information on this subject, the reader is invited to consider the recent survey by Deldjoo et al. [2020c]. Ultimately, recently, the public awareness of privacy issues has been steadily increasing. Preserving users' privacy in the recommendation learning process via building privacy-aware models (e.g., Anelli et al. [2019]) is gaining importance in real commercial systems.

Evaluation of multimedia recommender systems. The evaluation of the effectiveness of RS is an important research topic that goes beyond measuring the accuracy of the proposed recommendations [Kaminskas and Bridge 2017; Shani and Gunawardana 2011]. As outlined in Ge and Persia [2018], the evaluation of MMRS relies on the general evaluation procedures applied to RS; it could be worth investigating how the distinct media features and modalities involved in the recommendation process can impact on the effectiveness of the recommendations, not only from a purely accuracy-based perspective but considering also the user experience (e.g., perceived quality, usefulness, or satisfaction with the recommendations) and fairness of recommendation [Deldjoo et al. 2020a].

Considering the impact of user interfaces. The usefulness of recommendations in systems leveraging multimedia content is affected by the different design characteristics of the user interface (e.g., the format and presentation of the media content). The design of the user interface has a potential impact on the perceived quality of a multimedia recommender system [Cremonesi et al. 2017]. Algorithmic design dominates current research on multimedia recommender systems. However, several studies [Cremonesi et al. 2013] highlight a mismatch between algorithmic and user-perceived qualities and suggest the necessity for further study on novel user interfaces for recommender systems leveraging multimedia content [Amat et al. 2018].

Increasing transparency of multimedia recommender systems. Recommender systems are prominent applications of machine learning tools that directly interact with humans who are not necessarily experts in the field, e.g., a music educator or producer using a music recommendation service. Being able to offer interpretable output enables multimedia RS to explain their recommendations, increase algorithmic transparency, and in turn users' trust in and engagement with the system. These are important factors to motivate users to stay in and keep receiving recommendations, resulting in loyalty in the long term. Considering the recent advent and advances of deep learning models, which are currently widely adopted in research on MMRS, explainability and interpretability become even more important aspects, since neural network architectures per

se are often "black boxes" and hard or impossible to interpret. All the more, it is important to push research using deep neural networks to a position where we not only care about prediction accuracy but how users are able to make sense of why they have been recommended a particular media item.

To conclude, we believe that our survey will stimulate future research, bridging the fields of multimedia and recommender systems, by raising awareness of each other and featuring new topics to consider in future research endeavors.

APPENDIX: LIST OF ABBREVIATIONS USED IN THE SURVEY

Abbr.	Meaning	Abbr.	Meaning
AFF	attention fusion function	MLP	multi-layer perceptron
APC	automatic playlist continuation	MM	multimedia
BLF	block-level features	MMRS	multimedia recommender system
BPR	Bayesian personalized ranking	MRR	mean reciprocal rank
CA	context-aware	MSE	mean squared error
CBF	content-based filtering	PCA	principal components analysis
CBIR	content-based information retrieval	POI	point of interest
CCA	canonical correlation analysis	RMSE	root mean squared error
CF	collaborative filtering	RNN	recurrent neural network
CNN	convolutional neural network	RS	recommender syste
CS	cold-start	SCN	social curation networks
DBN	deep belief network	SIFT	scale invariant feature transform
DFT	discrete Fourier transform	SM	social media
DMF	deep matrix factorization	SN	social network
DNN	deep neural network	STFT	short-time Fourier transform
DWT	discrete Wavelet transform	SURF	speeded up robust features
HOG	histogram of oriented gradients	TF-IDF	term freq.–inverse document freq.
KB	knowledge base	UGC	user-generated content
KG	knowledge graph	UGV	user-generated video
LBP	local binary patterns	VBPR	Visual Bayesian personalized ranking
LBSN	location-based social network	VOD	video on-demand
LFM	latent factor model	WMF	weighted matrix factorization
LOD	linked open data	WPE	weighted prediction error
MF	matrix factorization	WS	warm-start
MFCC	Mel frequency cepstral coefficient	ZCR	zero crossing rate

REFERENCES

Charu C. Aggarwal. 2016a. Content-based recommender systems. In *Recommender Systems*. Springer, 139–166.

Charu C. Aggarwal. 2016b. Ensemble-based and hybrid recommender systems. In *Recommender Systems*. Springer, 199–224.

Taleb Alashkar, Songyao Jiang, and Yun Fu. 2017a. Rule-based facial makeup recommendation system. In *Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG'17)*. IEEE, 325–330.

Taleb Alashkar, Songyao Jiang, Shuyang Wang, and Yun Fu. 2017b. Examples-rules guided deep neural network for makeup recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 941–947.

Massimiliano Albanese, Angelo Chianese, Antonio d'Acierno, Vincenzo Moscato, and Antonio Picariello. 2010. A multimedia recommender integrating object features and user behavior. *Multimedia Tools Applic.* 50 (2010), 563–585.

Massimiliano Albanese, Antonio d'Acierno, Vincenzo Moscato, Fabio Persia, and Antonio Picariello. 2013. A multimedia recommender system. *ACM Trans. Internet Technol.* 13, 1 (2013).

J. Allen. 1977. Short term spectral analysis, synthesis, and modification by discrete Fourier transform. *IEEE Trans. Acoust. Speech Sig. Proc.* 25, 3 (June 1977), 235–238.

Fernando Amat, Ashok Chandrashekar, Tony Jebara, and Justin Basilico. 2018. Artwork personalization at Netflix. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 487–488.

Ivana Andjelkovic, Denis Parra, and John O'Donovan. 2019. Moodplay: Interactive music recommendation based on Artists' mood similarity. *Int. J. Hum.-comput. Stud.* 121 (2019), 142–159.

Vito Walter Anelli, Yashar Deldjoo, Tommaso Di Noia, and Antonio Ferrara. 2019. Towards effective device-aware federated learning. In *Proceedings of the International Conference of the Italian Association for Artificial Intelligence*. Springer, 477–491.

D. Azucar, D. Marengo, and M. Settanni. 2018. Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis. *Personal. Indiv. Dif.* 124 (2018), 150–159.

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 2011. *Modern Information Retrieval—The Concepts and Technology Behind Search* (2nd ed.). Addison-Wesley, Pearson, Harlow, England.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 2 (2019), 423–443.

Ilaria Bartolini, Vincenzo Moscato, Ruggero G. Pensa, Antonio Penta, Antonio Picariello, Carlo Sansone, and Maria Luisa Sapino. 2013. Recommending multimedia objects in cultural heritage applications. In *Proceedings of the International Conference on Image Analysis and Processing*. Springer, 257–267.

Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. In *Proceedings of the European Conference on Computer Vision*. Springer, 404–417.

Sergio Benini, Luca Canini, and Riccardo Leonardi. 2011. A connotative space for supporting movie affective recommendation. *IEEE Trans. Multimedia* 13, 6 (2011), 1356–1370.

Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*. 591–596.

Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. 2013. Recommender systems survey. *Knowledge-based Syst.* 46 (2013), 109–132.

Geoffray Bonnin and Dietmar Jannach. 2014. Automated generation of music playlists: Survey and experiments. *ACM Comput. Surv.* 47, 2 (Nov. 2014).

Steven Bourke, Kevin McCarthy, and Barry Smyth. 2011. The social camera: A case-study in contextual image recommendation. In *Proceedings of the 16th International Conference on Intelligent User Interfaces*. ACM, 13–22.

Sabri Boutemedjet and Djemel Ziou. 2006. A generative graphical model for collaborative filtering of visual content. In *Proceedings of the Industrial Conference on Data Mining*. Springer, 404–415.

Sabri Boutemedjet and Djemel Ziou. 2008. A graphical model for context-aware visual content recommendation. *IEEE Trans. Multimedia* 10, 1 (2008), 52–62.

Sabri Boutemedjet, Djemel Ziou, and Nizar Bouguila. 2008. Unsupervised feature selection for accurate recommendation of high-dimensional image data. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 177–184.

Jiajun Bu, Shulong Tan, Chun Chen, Can Wang, Hao Wu, Lijun Zhang, and Xiaofei He. 2010. Music recommendation by unified hypergraph: Combining social media information and music content. In *Proceedings of the ACM Multimedia Conference*. ACM, 391–400.

Robin Burke. 2002. Hybrid recommender systems: Survey and experiments. *User Model. User-adapt. Interact.* 12, 4 (2002), 331–370.

Luca Canini, Sergio Benini, and Riccardo Leonardi. 2013. Affective recommendation of movies based on selected connotative features. *IEEE Trans. Circ. Syst. Vid. Technol.* 23, 4 (2013), 636–647.

Erion Çano and Maurizio Morisio. 2019. Hybrid Recommender Systems: A Systematic Literature Review. *arxiv:cs.IR/1901.03888* (2019).

Fabio Celli, Elia Bruni, and Bruno Lepri. 2014. Automatic personality and interaction style recognition from Facebook profile pictures. In *Proceedings of the 22nd ACM International Conference on Multimedia (MM'14)*. ACM, New York, NY, 1101–1104.

S. Chen, J. L. Moore, D. Turnbull, and T. Joachims. 2012. Playlist prediction via metric embedding. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'12)*.

Xiaojie Chen, Pengpeng Zhao, Jiajie Xu, Zhixu Li, Lei Zhao, Yanchi Liu, Victor S. Sheng, and Zhiming Cui. 2018. Exploiting visual contents in posters and still frames for movie recommendation. *IEEE Access* 6 (2018), 68874–68881.

Heng-Yu Chi, Chun-Chieh Chen, Wen-Huang Cheng, and Ming-Syan Chen. 2016. UbiShop: Commercial item recommendation using visual part-based object representation. *Multimedia Tools Applic.* 75, 23 (2016), 16093–16115.

Wei-Ta Chu and Ya-Lun Tsai. 2017. A hybrid recommendation system considering visual information for predicting favorite restaurants. *World Wide Web* 20, 6 (2017), 1313–1331.

Kyung-Yong Chung. 2014. Effect of facial makeup style recommendation on visual sensibility. *Multimedia Tools Applic.* 71, 2 (2014), 843–853.

Paolo Cremonesi, Mehdi Elahi, and Franca Garzotto. 2017. User interface patterns in recommendation-empowered content intensive multimedia applications. *Multimedia Tools Applic.* 76, 4 (2017), 5275–5309.

Paolo Cremonesi, Franca Garzotto, and Roberto Turrin. 2013. User-centric vs. system-centric evaluation of recommender systems. In *Proceedings of the IFIP Conference on Human-Computer Interaction*. Springer, 334–351.

Guillem Cucurull, Pau Rodríguez, V. Oguz Yazici, Josep M. Gonfaus, F. Xavier Roca, and Jordi Gonzàlez. 2018. Deep inference of personality traits by integrating image and word use in social networks. *arXiv preprint arXiv:1802.06757* (2018).

Bin Cui, Anthony K. H. Tung, Ce Zhang, and Zhe Zhao. 2010. Multiple feature fusion for social media applications. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. ACM, 435–446.

Peng Cui, Zhiyu Wang, and Zhou Su. 2014. What videos are similar with you?: Learning a common attributed representation for video recommendation. In *Proceedings of the 22nd ACM International Conference on Multimedia*. ACM, 597–606.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys'19)*. 101–109.

Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. IEEE, 886–893.

Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. 2006. Studying aesthetics in photographic images using a computational approach. In *Proceedings of the European Conference on Computer Vision*. Springer, 288–301.

Marco de Gemmis, Pasquale Lops, Cataldo Musto, Fedelucio Narducci, and Giovanni Semeraro. 2015. Semantics-aware content-based recommender systems. In *Recommender Systems Handbook*. Springer, 119–159.

N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. 2011. Front-end factor analysis for speaker verification. *IEEE Trans. Aud. Speech, Lang. Proc.* 19, 4 (May 2011), 788–798.

Yashar Deldjoo, Vito Walter Anelli, Hamed Zamani, Alejandro Bellogin, and Tommaso Di Noia. 2020a. A flexible framework for evaluating user and item fairness in recommender systems. *User Modeling and User-Adapted Interaction* (2020).

Yashar Deldjoo, Mihai Gabriel Constantin, Hamid Eghbal-Zadeh, Bogdan Ionescu, Markus Schedl, and Paolo Cremonesi. 2018a. Audio-visual encoding of multimedia content for enhancing movie recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 455–459.

Yashar Deldjoo, Mihai Gabriel Constantin, Bogdan Ionescu, Markus Schedl, and Paolo Cremonesi. 2018b. MMTF-14k: A multifaceted movie trailer feature dataset for recommendation and retrieval. In *Proceedings of the 9th ACM Multimedia Systems Conference*. ACM, 450–455.

Yashar Deldjoo, Maurizio Ferrari Dacrema, Mihai Gabriel Constantin, Hamid Eghbal-Zadeh, Stefano Cereda, Markus Schedl, Bogdan Ionescu, and Paolo Cremonesi. 2019. Movie genome: Alleviating new item cold start in movie recommendation. *User Model. User-Adapt. Interact.* 29, 2 (2019), 291–343.

Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. 2020b. How dataset characteristics affect the robustness of collaborative recommendation models. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yashar Deldjoo, Mehdi Elahi, and Paolo Cremonesi. 2016a. Using visual features and latent factors for movie recommendation. In *Proceedings of the 3rd Workshop on New Trends in Content-Based Recommender Systems co-located with ACM Conference on Recommender Systems (RecSys'16), Boston, MA, USA, September 16, 2016*, Vol. 1673. CEUR-WS.org, 15–18. Retrieved from http://ceur-ws.org/Vol-1673/paper3.pdf.

Yashar Deldjoo, Mehdi Elahi, Paolo Cremonesi, Franca Garzotto, Pietro Piazzolla, and Massimo Quadrana. 2016b. Content-based video recommendation system based on stylistic visual features. *J. Data Seman.* 5, 2 (2016), 99–113.

Yashar Deldjoo, Mehdi Elahi, Massimo Quadrana, and Paolo Cremonesi. 2018c. Using visual features based on MPEG-7 and deep learning for movie recommendation. *Int. J. Multim. Inf. Retr.* 7, 4 (2018), 207–219. DOI: https://doi.org/10.1007/s13735-018-0155-1

Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. 2020c. Adversarial machine learning in recommender systems: State of the art and challenges. *CoRR* abs/2005.10322 (2020).

Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2018d. Content-based multimedia recommendation systems: Definition and application domains. In *Proceedings of the 9th Italian Information Retrieval Workshop, Rome, Italy, May, 28-30, 2018*, Vol. 2140. CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-2140/paper15.pdf.

Zhengyu Deng, Jitao Sang, and Changsheng Xu. 2013. Personalized video recommendation based on cross-platform user modeling. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'13)*. IEEE, 1–6.

Tommaso Di Noia, Daniele Malitesta, and Felice Antonio Merra. 2020. TAaMR: Targeted adversarial attack against multimedia recommender systems. In *Proceedings of the 3rd International Workshop on Dependable and Secure Machine Learning (DSML'20)*. IEEE.

Xingzhong Du, Hongzhi Yin, Ling Chen, Yang Wang, Yi Yang, and Xiaofang Zhou. 2020. Personalized video recommendation using rich contents from videos. *IEEE Trans. Knowl. Data Eng.* 32, 3 (2020), 492–505.

Michael D. Ekstrand, John T. Riedl, Joseph A. Konstan et al. 2011. Collaborative filtering recommender systems. *Found. Trends® Hum.-comput. Interact.* 4, 2 (2011), 81–173.

David Elsweiler, Christoph Trattner, and Morgan Harvey. 2017. Exploiting food choice biases for healthier recipe recommendation. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 575–584.

Aleksandr Farseev, Liqiang Nie, Mohammad Akbari, and Tat-Seng Chua. 2015. Harvesting multiple sources for user profile learning: A big data study. In *Proceedings of the 5th ACM International Conference on Multimedia Retrieval*. ACM, 235–242.

Aleksandr Farseev, Ivan Samborskii, Andrey Filchenkov, and Tat-Seng Chua. 2017. Cross-domain recommendation via clustering on multi-layer graphs. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 195–204.

Bruce Ferwerda and Marko Tkalcic. 2018. Predicting users' personality from Instagram pictures: Using visual and/or content features? In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization (UMAP'18)*. ACM, New York, NY, 157–161.

Bruce Ferwerda, Marko Tkalcic, and Markus Schedl. 2017. Personality traits and music genres: What do people prefer to listen to? In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP'17)*. ACM, New York, NY, 285–288.

Zeno Gantner, Lucas Drumond, Christoph Freudenthaler, Steffen Rendle, and Lars Schmidt-Thieme. 2010. Learning attribute-to-feature mappings for cold-start recommendations. In *Proceedings of the IEEE 10th International Conference on Data Mining (ICDM'10)*. IEEE, 176–185.

Mouzhi Ge and Fabio Persia. 2018. Evaluation in multimedia recommender systems: A practical guide. In *Proceedings of the 12th IEEE International Conference on Semantic Computing (ICSC'18)*. 294–297.

Xue Geng, Hanwang Zhang, Jingwen Bian, and Tat-Seng Chua. 2015. Learning image and user features for recommendation in social networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'15)*. 4274–4282.

Jennifer Golbeck and Eric Norris. 2013. Personality, movie preferences, and recommendations. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, New York, NY, 1414–1415.

Samuel D. Gosling, Peter J. Rentfrow, and William B. Swann Jr. 2003. A very brief measure of the big-five personality domains. *J. Res. Personal.* 37, 6 (2003), 504–528.

Xiaoling Gu, Lidan Shou, Pai Peng, Ke Chen, Sai Wu, and Gang Chen. 2016. iGlasses: A novel recommendation system for best-fit glasses. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1109–1112.

Sharath Chandra Guntuku, Lin Qiu, Sujoy Roy, Weisi Lin, and Vinit Jakhetiya. 2015a. Do others perceive you as you want them to?: Modeling personality based on selfies. In *Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia*. ACM, 21–26.

Sharath Chandra Guntuku, Sujoy Roy, and Lin Weisi. 2015b. Personality modeling based image recommendation. In *Proceedings of the International Conference on Multimedia Modeling*. Springer, 171–182.

Sharath Chandra Guntuku, Michael James Scott, Gheorghita Ghinea, and Weisi Lin. 2016. Personality, culture, and system factors-impact on affective response to multimedia. *arXiv preprint arXiv:1606.06873* (2016).

Mario Haim, Andreas Graefe, and Hans-Bernd Brosius. 2018. Burst of the filter bubble? Effects of personalization on the diversity of Google News. *Dig. J.* 6, 3 (2018), 330–343.

Ruining He and Julian McAuley. 2016a. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the WWW Conference*. 507–517.

Ruining He and Julian McAuley. 2016b. VBPR: Visual Bayesian personalized ranking from implicit feedback. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*. 144–150.

Luis Herranz, Weiqing Min, and Shuqiang Jiang. 2018. Food recognition and recipe analysis: Integrating visual content, context, and external knowledge. *arXiv preprint arXiv:1801.07239* (2018).

Prajakta A. Holey and S. S. Prabhune. 2014. Review of content-based recommendation system. *Int. J. Sci. Eng. Technol. Res.* 3, 4 (2014).

Min Hou, Le Wu, Enhong Chen, Zhi Li, Vincent W. Zheng, and Qi Liu. 2019. Explainable fashion recommendation: A semantic attribute region guided approach. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI'19)*. 4681–4688.

Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM'08)*. 263–272.

Mark J. Huiskes and Michael S. Lew. 2008. The MIR flickr retrieval evaluation. In *Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval (MIR'08)*. 39–43.

Oliver John and Sanjay Srivastava. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of Personality: Theory and Research* (2nd ed.), Lawrence A. Pervin and Oliver P. John (Eds.). Guilford Press, New York, 102–138.

Oliver P. John, Eileen M. Donahue, and Robert L. Kentle. 1991. The big five inventory. *Journal of Personality and Social Psychology* (1991).

Marius Kaminskas and Derek Bridge. 2017. Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. *Trans. Internet Inf. Syst.* 7, 1 (2017), 2:1–2:42.

Marius Kaminskas, Francesco Ricci, and Markus Schedl. 2013. Location-aware music recommendation using auto-tagging and hybrid matching. In *Proceedings of the ACM Conference on Recommender Systems*. ACM, 17–24.

Peter Knees and Markus Schedl. 2013. A survey of music similarity and recommendation from music context data. *ACM Trans. Multimedia Comput. Commun. Applic.* 10, 1 (2013).

Yehuda Koren and Robert Bell. 2015. Advances in collaborative filtering. In *Recommender Systems Handbook*. Springer, 77–118.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (Aug. 2009), 30–37.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 1097–1105.

Xingchen Li, Xiang Wang, Xiangnan He, Long Chen, Jun Xiao, and Tat-Seng Chua. 2020. Hierarchical fashion graph network for personalized outfit recommendation. *CoRR* abs/2005.12566 (2020).

Dawen Liang, Minshu Zhan, and Daniel P. W. Ellis. 2015. Content-aware collaborative music recommendation using pre-trained neural networks. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR'15)*. 295–301.

Jimmy Lin. 2019. The neural hype and comparisons against weak baselines. *SIGIR Forum* 52, 2 (Jan. 2019), 40–51.

Q. Lin, Y. Niu, Y. Zhu, H. Lu, K. Z. Mushonga, and Z. Niu. 2018. Heterogeneous knowledge-based attentive neural networks for short-term music recommendations. *IEEE Access* 6 (2018), 58990–59000.

Yusan Lin, Maryam Moosaei, and Hao Yang. 2020. OutfitNet: Fashion outfit recommendation with attention-based multiple instance learning. In *Proceedings of the WWW Conference (WWW'20)*. ACM/IW3C2, 77–87.

Zijia Lin, Guiguang Ding, and Jianmin Wang. 2011. Image annotation based on recommendation model. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1097–1098.

Jing Liu, Zechao Li, Jinhui Tang, Yu Jiang, and Hanqing Lu. 2014. Personalized geo-specific tag recommendation for photos on social websites. *IEEE Trans. Multimedia* 16, 3 (2014), 588–600.

Beth Logan. 2000. Mel frequency cepstral coefficients for music modeling. In *Proceedings of the International Symposium on Music Information Retrieval (ISMIR'00)*.

David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60, 2 (2004), 91–110.

Hangzai Luo, Jianping Fan, Daniel A. Keim, and Shin'ichi Satoh. 2009. Personalized news video recommendation. In *Proceedings of the International Conference on Multimedia Modeling*. Springer, 459–471.

Jingwei Ma, Guang Li, Mingyang Zhong, Xin Zhao, Lei Zhu, and Xue Li. 2018. LGA: Latent genre aware micro-video recommendation on social media. *Multimedia Tools Applic.* 77, 3 (2018), 2991–3008.

Anand Mahadevan, Jason Freeman, Brian Magerko, and Juan Carlos Martinez. 2015. EarSketch: Teaching computational music remixing in an online web audio–based learning environment. In *Proceedings of the Web Audio Conference*.

Bangalore S. Manjunath and Wei-Ying Ma. 1996. Texture features for browsing and retrieval of image data. *IEEE Trans. Pattern Anal. Mach. Intell.* 18, 8 (1996), 837–842.

Richard E. Mayer. 2005. *The Cambridge Handbook of Multimedia Learning, 1st Edition*. Cambridge University Press. Retrieved from http://www.worldcat.org/oclc/57526976.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 43–52.

Brian McFee, Luke Barrington, and Gert Lanckriet. 2012. Learning content similarity for music recommendation. *IEEE Trans. Aud. Speech, Lang. Proc.* 20, 8 (2012), 2207–2218.

Brian McFee and Gert Lanckriet. 2011. The natural language of playlists. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR'11)*.

Tao Mei, Bo Yang, Xian-Sheng Hua, and Shipeng Li. 2011. Contextual video recommendation by multimodal relevance and user feedback. *ACM Trans. Inf. Syst.* 29, 2 (2011), 10.

Tao Mei, Bo Yang, Xian-Sheng Hua, Linjun Yang, Shi-Qiang Yang, and Shipeng Li. 2007. VideoReach: An online video recommendation system. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 767–768.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations (ICLR'13)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS'13)*, Vol. 2. Curran Associates Inc., 3111–3119.

Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'12)*. IEEE, 2408–2415.

Radford M. Neal. 2000. Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Stat.* 9, 2 (2000), 249–265.

Robert Neumayer and Andreas Rauber. 2007. Integration of text and audio features for genre classification in music information retrieval. In *Proceedings of the European Conference on Information Retrieval*. Springer, 724–727.

Vinh-Tiep Nguyen, Khanh-Duy Le, Minh-Triet Tran, and Morten Fjeld. 2016. NowAndThen: A social network-based photo recommendation tool supporting reminiscence. In *Proceedings of the 15th International Conference on Mobile and Ubiquitous Multimedia*. ACM, 159–168.

Wei Niu, James Caverlee, and Haokai Lu. 2018. Neural personalized ranking for image recommendation. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. ACM, 423–431.

Timo Ojala, Matti Pietikainen, and Topi Maenpaa. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 7 (2002), 971–987.

Sergio Oramas, Oriol Nieto, Mohamed Sordo, and Xavier Serra. 2017. A deep multimodal approach for cold-start music recommendation. In *Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems (DLRS'17)*. ACM, New York, NY, 32–37.

Sergio Oramas, Vito Claudio Ostuni, Tommaso Di Noia, Xavier Serra, and Eugenio Di Sciascio. 2016. Sound and music recommendation with knowledge graphs. *ACM Trans. Intell. Syst. Technol.* 8, 2 (Oct. 2016).

Claudia Orellana-Rodriguez, Ernesto Diaz-Aviles, and Wolfgang Nejdl. 2015. Mining affective context in short films for emotion-aware recommendation. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media (HT'15)*. ACM, New York, NY, 185–194.

Vito Claudio Ostuni, Tommaso Di Noia, Eugenio Di Sciascio, Sergio Oramas, and Xavier Serra. 2015. A semantic hybrid approach for sound recommendation. In *Proceedings of the WWW Conference*. 85–86.

Deuk Hee Park, Hyea Kyeong Kim, Il Young Choi, and Jae Kyeong Kim. 2012. A literature review and classification of recommender systems research. *Exp. Syst. Applic.* 39, 11 (2012), 10059–10072.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1532–1543.

Florent Perronnin and Christopher Dance. 2007. Fisher kernels on visual vocabularies for image categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*. IEEE, 1–8.

Ladislav Peska and Hana Trojanova. 2017. Towards recommender systems for police photo lineup. *arXiv preprint arXiv:1707.01389* (2017).

Gabriele Prato, Federico Sallemi, Paolo Cremonesi, Mario Scriminaci, Stefan Gudmundsson, and Silvio Palumbo. 2020. Outfit completion and clothes recommendation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'20)*. ACM, 1–7.

Pearl Pu and Li Chen. 2007. Trust-inspiring explanation interfaces for recommender systems. *Knowl.-based Syst.* 20, 6 (2007), 542–556.

Xueming Qian, He Feng, Guoshuai Zhao, and Tao Mei. 2014. Personalized recommendation combining user interest and social circle. *IEEE Trans. Knowl. Data Eng.* 26, 7 (2014), 1763–1777.

Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. 2018. Sequence-aware recommender systems. *Comput. Surv.* 51, 4 (July 2018).

Amir Hossein Nabizadeh Rafsanjani, Naomie Salim, Atae Rezaei Aghdam, and Karamollah Bagheri Fard. 2013. Recommendation systems: A review. *Int. J. Comput. Eng. Res.* 3, 5 (2013), 47–52.

Yogesh Singh Rawat and Mohan S. Kankanhalli. 2017. ClickSmart: A context-aware viewpoint recommendation system for mobile photography. *IEEE Trans. Circ. Syst. Video Technol.* 27, 1 (2017), 149–158.

Peter Rentfrow, Lewis R. Goldberg, and Ran Zilca. 2011. Listening, watching, and reading: The structure and correlates of entertainment preferences. *J. Personal.* 79 (Apr. 2011), 223–258.

Sujoy Roy and Sharat Chandra Guntuku. 2016. Latent factor representations for cold-start video recommendation. In *Proceedings of the ACM Conference on Recommender Systems*. ACM, 99–106.

Noveen Sachdeva, Kartik Gupta, and Vikram Pudi. 2018. Attentive neural architecture incorporating song features for music recommendation. In *Proceedings of the ACM Conference on Recommender Systems (RecSys'18)*. ACM, New York, NY, 417–421.

Markus Schedl. 2019. Deep learning in music recommender systems. *Frontiers in Applied Mathematics and Statistics* 5 (2019), 44 pages.

Markus Schedl, Arthur Flexer, and Julián Urbano. 2013. The neglected user in music information retrieval research. *J. Intell. Inf. Syst.* 41, 3 (Dec. 2013), 523–539.

Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. 2018. Current challenges and visions in music recommender systems research. *Int. J. Multimedia Inf. Retr.* 7, 2 (2018), 95–116.

Jan Schlüter. 2016. Learning to pinpoint singing voice from weakly labeled examples. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR'16)*.

Klaus Seyerlehner, Gerhard Widmer, and Tim Pohle. 2010. Fusing block-level features for music similarity estimation. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx'10)*.

Dandan Sha, Daling Wang, Xiangmin Zhou, Shi Feng, Yifei Zhang, and Ge Yu. 2016. An approach for clothing recommendation based on multiple image attributes. In *Proceedings of the International Conference on Web-age Information Management*. Springer, 272–285.

Rajiv Ratn Shah, Yi Yu, and Roger Zimmermann. 2014. ADVISOR: Personalized video soundtrack recommendation by late fusion with heuristic rankings. In *Proceedings of the ACM International Conference on Multimedia (MM'14)*. 607–616.

Guy Shani and Asela Gunawardana. 2011. Evaluating recommendation systems. *Recommender Systems Handbook*. Springer, 257–297.

Bo Shao, Dingding Wang, Tao Li, and Mitsunori Ogihara. 2009. Music recommendation based on acoustic features and user access patterns. *IEEE Trans. Aud. Speech, Lang. Proc.* 17, 8 (2009), 1602–1611.

Yue Shi, Alexandros Karatzoglou, Linas Baltrunas, Martha Larson, Nuria Oliver, and Alan Hanjalic. 2012. CLiMF: Learning to maximize reciprocal rank with collaborative less-is-more filtering. In *Proceedings of the ACM Conference on Recommender Systems*. ACM, 139–146.

Yue Shi, Martha Larson, and Alan Hanjalic. 2014. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Comput. Surv.* 47, 1 (2014), 3.

Marcin Skowron, Bruce Ferwerda, Marko Tkalčič, and Markus Schedl. 2016. Fusing social media cues: Personality prediction from Twitter and Instagram. In *Proceedings of the WWW Conference*.

Jason Smith, Dillon Weeks, Mikhail Jacob, Jason Freeman, and Brian Magerko. 2019. Towards a hybrid recommendation system for a sound library. In *Joint Proceedings of the ACM IUI Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces (ACM-IUI'19)*.

Adrian Stanciulescu. 2008. *A Methodology for Developing Multimodal User Interfaces of Information Systems*. Ph.D. Dissertation. Catholic University of Louvain, Louvain-la-Neuve, Belgium. Retrieved from http://hdl.handle.net/2078.1/12738.

Guang-Lu Sun, Zhi-Qi Cheng, Xiao Wu, and Qiang Peng. 2018. Personalized clothing recommendation combining user social circle and fashion style consistency. *Multimedia Tools Applic.* 77, 14 (2018), 17731–17754.

Jinhui Tang, Xiaoyu Du, Xiangnan He, Fajie Yuan, Qi Tian, and Tat-Seng Chua. 2020. Adversarial training towards robust multimedia recommender system. *IEEE Trans. Knowl. Data Eng.* 32, 5 (2020), 855–867. DOI: https://doi.org/10.1109/TKDE.2019.2893638

Marko Tkalcic and Li Chen. 2015. *Personality and Recommender Systems*. Springer US, Boston, MA, 715–739.

Christoph Trattner, Dominik Moesslang, and David Elsweiler. 2018. On the predictability of the popularity of online recipes. *EPJ Data Sci.* 7, 1 (2018), 20.

Aäron van den Oord, Sander Dieleman, and Benjamin Schrauwen. 2013. Deep content-based music recommendation. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems (NIPS'13)*, Christopher Burges, Léon Bottou, Max Welling, Zoubin Ghahramani, and Kilian Weinberger (Eds.). Curran Associates, Inc., 2643–2651.

Andreu Vall, Matthias Dorfer, Hamid Eghbal-zadeh, Markus Schedl, Keki Burjorjee, and Gerhard Widmer. 2019. Feature-combination hybrid recommender systems for automated music playlist continuation. *User Model. User-Adapt. Interact.* 29, 2 (2019), 527–572. DOI: https://doi.org/10.1007/s11257-018-9215-8

Shangfei Wang and Qiang Ji. 2015. Video affective content analysis: A survey of state-of-the-art methods. *IEEE Trans. Affect. Comput.* 6, 4 (2015), 410–430.

Suhang Wang, Yilin Wang, Jiliang Tang, Kai Shu, Suhas Ranganath, and Huan Liu. 2017. What your images reveal: Exploiting visual contents for point-of-interest recommendation. In *Proceedings of the WWW Conference*. 391–400.

Xinxi Wang and Ye Wang. 2014. Improving content-based and hybrid music recommendation using deep learning. In *Proceedings of the 22nd ACM International Conference on Multimedia*. ACM, 627–636.

Zhangyang Wang, Shiyu Chang, Florin Dolcos, Diane Beck, Ding Liu, and Thomas S. Huang. 2016. Brain-inspired deep networks for image aesthetics assessment. *arXiv preprint arXiv:1601.04155* (2016).

Kangning Wei, Jinghua Huang, and Shaohong Fu. 2007. A survey of e-commerce recommender systems. In *Proceedings of the International Conference on Service Systems and Service Management*. IEEE, 1–5.

Jiqing Wen, James She, Xiaopeng Li, and Hui Mao. 2018. Visual background recommendation for dance performances using deep matrix factorization. *ACM Trans. Multimedia Comput. Commun. Applic.* 14, 1 (2018), 11:1–11:19.

Stina Westman and Pirkko Oittinen. 2006. Image retrieval by end-users and intermediaries in a journalistic work context. In *Proceedings of the 1st International Conference on Information Interaction in Context*. ACM, 102–110.

Chun-Che Wu, Tao Mei, Winston H. Hsu, and Yong Rui. 2014. Learning to personalize trending image search suggestion. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 727–736.

Feng Xia, Nana Yaw Asabere, Ahmedin Mohammed Ahmed, Jing Li, and Xiangjie Kong. 2013. Mobile multimedia recommendation in smart communities: A survey. *IEEE Access* 1 (2013), 606–624.

Zhou Xing, Marzieh Parandehgheibi, Fei Xiao, Nilesh Kulkarni, and Chris Pouliot. 2016. Content-based recommendation for podcast audio-items using natural language processing techniques. In *Proceedings of the IEEE International Conference on Big Data (Big Data'16)*. IEEE, 2378–2383.

Bo Yang, Tao Mei, Xian-Sheng Hua, Linjun Yang, Shi-Qiang Yang, and Mingjing Li. 2007. Online video recommendation based on multimodal fusion and relevance feedback. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*. ACM, 73–80.

Longqi Yang, Cheng-Kang Hsieh, Hongjian Yang, John P. Pollak, Nicola Dell, Serge Belongie, Curtis Cole, and Deborah Estrin. 2017. Yum-me: A personalized nutrient-based meal recommender system. *ACM Trans. Inf. Syst.* 36, 1 (2017), 7.

Longqi Yang, Michael Sobolev, Christina Tsangouri, and Deborah Estrin. 2018. Understanding user interactions with podcast recommendations delivered via voice. In *Proceedings of the ACM Conference on Recommender Systems*. ACM, 190–194.

Kazuyoshi Yoshii, Masataka Goto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. 2007. Improving efficiency and scalability of model-based music recommender system based on incremental training. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR'07)*.

Kazuyoshi Yoshii, Masataka Goto, Kazuhiro Komatani, Tetsuya Ogata, and Hiroshi G. Okuno. 2008. An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model. *IEEE Trans. Aud. Speech Lang. Proc.* 16, 2 (2008), 435–447.

Dongfei Yu, Xinmei Tian, Tao Mei, and Yong Rui. 2015. On the selection of trending image from the web. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'15)*. IEEE, 1–6.

Wenhui Yu, Huidi Zhang, Xiangnan He, Xu Chen, Li Xiong, and Zheng Qin. 2018. Aesthetic-based clothing recommendation. In *Proceedings of the WWW Conference*. 649–658.

Jianbo Yuan, Walid Shalaby, Mohammed Korayem, David Lin, Khalifeh AlJadda, and Jiebo Luo. 2016. Solving cold-start problem in large-scale recommendation engines: A deep learning approach. In *Proceedings of the IEEE International Conference on Big Data (Big Data'16)*. IEEE, 1901–1910.

Hamed Zamani, Markus Schedl, Paul Lamere, and Ching-Wei Chen. 2018. An analysis of approaches taken in the ACM recsys challenge 2018 for automatic music playlist continuation. *CoRR* arXiv:1810.01520 (2018).

Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM Comput. Surv.* 52, 1 (2019), 1–38.

Yongfeng Zhang and Xu Chen. 2020. Explainable recommendation: A survey and new perspectives. *Found. Trends Inf. Retr.* 14, 1 (2020), 1–101. Retrieved from DOI: https://doi.org/10.1561/1500000066.

Xiaojian Zhao, Huanbo Luan, Junjie Cai, Jin Yuan, Xiaoming Chen, and Zhoujun Li. 2012a. Personalized video recommendation based on viewing history with the study on YouTube. In *Proceedings of the 4th International Conference on Internet Multimedia Computing and Service*. ACM, 161–165.

Xiaojian Zhao, Jin Yuan, Richang Hong, Meng Wang, Zhoujun Li, and Tat-Seng Chua. 2012b. On video recommendation over social network. In *Proceedings of the International Conference on Multimedia Modeling*. Springer, 149–160.

Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proc. Nat. Acad. Sci.* 107, 10 (2010), 4511–4515.

Qiusha Zhu, Mei-Ling Shyu, and Haohong Wang. 2013. Videotopic: Content-based video recommendation using a topic model. In *Proceedings of the IEEE International Symposium on Multimedia (ISM'13)*. IEEE, 219–222.

Received September 2019; revised June 2020; accepted June 2020